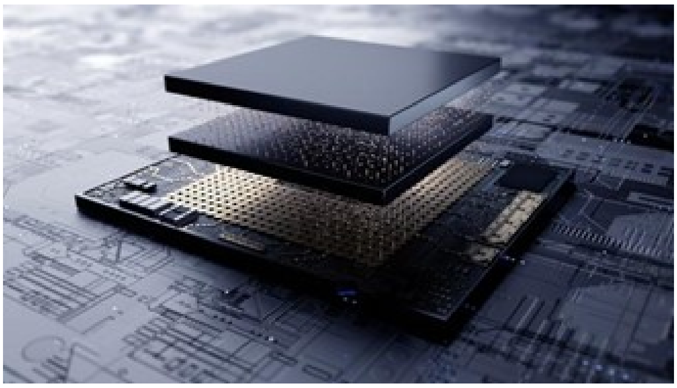


Ultra-Low Power Performance Sensor for CMOS Memory Cells

Nuno Miguel da Rocha Calisto

Thesis to obtain the Master of Science Degree in

Aerospace Engineering

Supervisors: Prof. Jorge Filipe Leal Costa Semião
Prof. Marcelino Bicho dos Santos

Examination Committee

Chairperson: Prof. Paulo Jorge Coelho Ramalho Oliveira
Supervisor: Prof. Marcelino Bicho dos Santos
Member of the Committee: Prof. Jorge Manuel Dos Santos Ribeiro Fernandes

December 2022

Acknowledgments

I would like to thank my supervisors, Prof. Jorge Semião and Prof. Marcelino Santos, for all the time available in meetings, for clarifying doubts and extremely useful suggestions for the preparation of this master's thesis.

I would also like to thank Prof. João Silva for giving me access to a PC of INESC, for providing the necessary technology for the realization of this dissertation and clarification of doubts.

To Prof. Fernando Lau, coordinator of the Master's degree in Aerospace Engineering at IST, thank you for encouraging me all these years to study and improve my knowledge.

To INESC, thank you for all the resources available for the realization of this dissertation.

To my classmates and friends, thank you for your encouragement and help over the years.

Finally, I would like to thank my parents for all the financial and moral support because without them the realization of this dissertation would not be possible.

Resumo

Com o avanço da tecnologia, cada vez mais a nossa sociedade utiliza dispositivos eletrónicos que contêm memórias para guardar as suas instruções. Nos circuitos integrados, as memórias do tipo *Complementary Metal Oxide Semiconductor* (CMOS) são as mais utilizadas e, com o passar dos tempos, o seu tamanho reduzido pode provocar problemas de performance e de fiabilidade. Estes problemas podem ser provocados por efeitos de envelhecimento, tais como o BTI (*Bias Thermal Instability*), o TDDB (*Time Dependent Dielectric Breakdown*), o HCI (*Hot Carrier Injection*) e o EM (*Electromigration*), que vão deteriorando os parâmetros físicos dos transístores MOSFET, mudando as suas propriedades elétricas.

Os efeitos BTI contêm dois tipos de efeitos de envelhecimento: o efeito PBTI (*Positive* BTI), que afeta mais os transístores NMOS, e o efeito NBTI (*Negative* BTI), que afeta mais os transístores PMOS e são mais visíveis para nanotecnologias até 32 nanometros. Para além dos efeitos de envelhecimento (*Aging* - A), existem ainda outras variações no desempenho que podem colocar em causa o bom funcionamento dos circuitos, como as variações de processo (P), tensão (V) e temperatura (T), que todas juntas formam os efeitos PVTA.

Considerando as memórias RAM (*Random Access Memory*), em particular as memórias SRAM (*Static Random Access Memory*) e as memórias DRAM (*Dynamic Random Access Memory*), estas podem ficar expostas ao envelhecimento dos seus componentes, provocando um decréscimo na sua performance, resultando em transições mais lentas, que por sua vez irão provocar leituras e escritas mais lentas, que podem dar origem à ocorrência de erros durante essas operações. Portanto, o envelhecimento das memórias CMOS traduz-se na ocorrência de erros nas memórias ao longo do tempo, o que é indesejável, especialmente em sistemas críticos. Torna-se assim necessário monitorizar os erros de uma memória através de sensores.

Outra questão crucial para aplicações IoT (*Internet-of-Things*) é a gestão de energia. Uma grande variedade de sensores inteligentes, geralmente operados por bateria, requer alta eficiência energética. Isto visa a busca por microcontroladores de potência ultrabaixa e memórias de baixa potência. Para isso, uma variável chave é o valor mínimo da tensão da fonte de alimentação, V_{DD}, que pode garantir a retenção segura dos dados e o acesso aos dados (operações de leitura/escrita). Usando uma unidade de gestão de energia flexível (PMU – *Power Management Unit*), pode ser recompensador realizar o dimensionamento dinâmico de tensão e frequência (DVFS – *Dynamic Voltage and Frequency Scaling*) para alimentar matrizes de memória com V_{DD} mínimo durante o acesso à memória e retenção de dados.

Já foram realizados alguns trabalhos sobre sensores que permitem monitorizar os erros de uma memória, como é o caso do sensor OCAS (On-Chip Aging Sensor) que detecta envelhecimento numa memória SRAM provocado pelo envelhecimento por NBTI. No entanto, este sensor apresenta algumas limitações, pois não pode ser aplicado em memórias DRAM e não contempla o efeito PBTI. Outra solução apresentada anteriormente é o sensor de performance para uma memória SRAM realizado por Hugo Santos [1], que demonstra alguma evolução em relação ao sensor OCAS, mas ainda contém limitações, como é o caso de ser bastante dependente do sincronismo com a memória e não permitir qualquer tipo de calibração do sistema ao longo do seu funcionamento. Com o objetivo de ultrapassar as limitações do sensor anterior, foi apresentado por Luís Santos o *Scout Memory Sensor* [2], que permite o seu uso em memórias SRAM e DRAM e também permite, ao projetista, calibrar e mudar a sensibilidade do sensor, tornando esta solução mais versátil e robusta. No entanto, o *Scout Memory Sensor* não é consistente e coerente para um regime *subthreshold*, não garantindo a sinalização dos erros quando as tensões de alimentação são muito baixas. Torna-se assim necessário encontrar uma solução alternativa para este sensor que funcione a baixas tensões de alimentação.

Esta dissertação tem como objetivo apresentar um novo sensor (*Ultra-Low Power Performance Sensor for CMOS Memory Cells*), para superar os problemas detetados no *Scout Memory Sensor*. Este sensor é compatível com vários tipos de memória e arquitecturas (SRAM e DRAM) e é um sensor de performance que deteta a degradação provocada pelas variações PVTA com baixo consumo de potência, utilizando técnicas de DVFS, permitindo assim o seu uso para tensões de alimentação (V_{DD}) menores, com o objetivo de poupar energia. Este *Ultra-Low Power Performance Sensor for CMOS Memory Cells* representa uma novidade em relação aos sensores anteriormente propostos, por isso ainda não foi testado em circuitos reais.

A arquitetura do *Ultra-Low Power Performance Sensor* é composta pelo bloco *transition detector*, por dois *delay elements* e por um bloco *flip-flop*. Este sensor sinaliza a degradação dos sinais durante as operações de Leitura/Escrita numa célula de memória devido a variações PVTA. Com estas variações, os atrasos nas transições ocorridas na *bit line* e os atrasos de propagação dos sinais pelas portas lógias aumentam e, quando o somatório total dos atrasos das portas de um sensor ligado a uma *bit line* ultrapassa um período de relógio, o sensor sinaliza um erro, indicando que o funcionamento normal da memória está na iminência de falhar. Na realidade, o sensor deteta uma transição da *bit line* e gera um impulso proporcional ao tempo de transição. Devido à presença de elementos de atraso no sensor, é adicionado um atraso ao impulso, que permite capturar o impulso atrasado por um *flip-flop* e sinalizar o erro preditivo. Caso o impulso atrasado não seja capturado pelo *flip-flop*, não há sinalização.

Como já foi referido anteriormente, este sensor é compatível com a utilização de técnicas de DVFS, pois pode ser utilizado com baixas tensões de alimentação. O objetivo de utilização do sensor com esta técnica é a de corresponder cada valor de VDD com uma frequência diferente do relógio, de modo a que o sensor funcione sempre na iminência de um erro, permitindo operar o circuito na máxima performance, ou mínima energia, garantindo que a margem de segurança se mantenha semelhante para todos os níveis de V_{DD}.

Este sensor pode ainda ser utilizado internamente na memória, como sensor local (monitorizando as células reais de memória), ou externamente, como sensor global, caso seja colocado na monitorizar uma célula de memória fictícia.

Palavras-chave: SRAM, DRAM, Sensor de performance, memórias, CMOS

Abstract

With the advancement of technology, our society increasingly uses electronic devices that contain memories to keep their instructions. In integrated circuits, Complementary Metal Oxide Semiconductor (CMOS) memory is the most widely used and over time their reduced size can cause performance and reliability problems. These problems can be caused by aging effects, such as BTI (Bias Thermal Instability), TDDB (Time Dependent Dielectric Breakdown), HCI (Hot Carrier Injection) and EM (Electromigration), which deteriorate the physical parameters of MOSFET transistors, changing their electrical properties.

BTI effects contain two types of aging effects: PBTI (Positive BTI) effect, which affects NMOS transistors more and the NBTI (Negative BTI) effect, which affects PMOS transistors more and is more visible for nanotechnologies up to 32 nanometers. In addition to the effects of aging (A), there are also other variations in performance that can call into question the proper functioning of the circuits, such as process variations (P), voltage (V) and temperature (T), which all together form PVTA effects.

Considering Random Access Memory (RAM), Static Random Access Memory (SRAM) and Dynamic Random Access Memory (DRAM) memories, these can be exposed to aging of their components, causing a decrease in their performance, resulting in slower transitions, which in turn will cause slower reads and writes, which can lead to errors during these operations. Therefore, the aging of CMOS memories translates into the occurrence of errors in memories over time, which is undesirable, especially in critical systems. It is therefore necessary to monitor the errors of a memory through sensors.

Another crucial issue for IoT applications is energy management. A wide variety of smart sensors, usually battery operated, require high energy efficiency. This is aimed at searching for ultra-low power microcontrollers and low-power memories. For this, a key variable is the minimum power supply voltage value, V_{DD}, which can ensure secure data retention and access to data (read/write operations). Using a flexible power management unit (PMU), it can be rewarding to perform dynamic voltage and frequency sizing (DVFS) to power memory matrices with minimal V_{DD} during memory access and data retention.

Some work has already been done in the search for sensors that allow monitoring the errors of a memory, such as the OCAS sensor (On-Chip Aging Sensor), which detects aging in an SRAM memory caused by aging by NBTI. However, this sensor has some limitations, as it cannot be applied to DRAM memories and does not contemplate the PBTI effect. Another solution presented earlier is the performance sensor for a SRAM memory performed by Hugo Santos that demonstrates some evolution in relation to the OCAS sensor, but still contains limitations, as is the case of being quite dependent on the synchronism with the memory and not allowing any type of calibration of the system throughout its operation. In order to overcome the limitations of the previous sensor, the Scout Memory Sensor was presented by Luis Santos, which allows its use in SRAM and DRAM memories and also allows the designer to calibrate and change the sensitivity of the sensor, making this solution more versatile and robust. However, the Scout Memory Sensor is not consistent and coherent for a subthreshold regime, not ensuring error signaling when supply voltages are too low. It is therefore necessary to find a alternative solution for this sensor that work at low supply voltages.

This dissertation aims to present a new sensor (Ultra-Low Power Performance Sensor for Memory Cells), to overcome the problems detected in the Scout Memory Sensor. This sensor is compatible with various types of memory and architectures (SRAM and DRAM) and is a performance sensor that detects the degradation caused by PVTA variations with low power consumption, using DVFS (Dynamic Voltage and Frequency Scaling) techniques, thus allowing its use for lower supply voltages (V_{DD}) in order to save energy. This Ultra-Low Power Performance Sensor for Memory Cells is a novelty compared to the previously proposed sensors, so it has not yet been tested on real circuits.

The architecture of the Ultra-Low Power Performance Sensor is composed of a transition detector block, two delay elements and a flip-flop. This sensor detects the degradation of signals during Read/Write operations in a memory cell due to PVTA variations. With these variations, the delays in the transitions that occur in the bit line and the propagation delays of the signals through the logic gates increase and, when the total sum of the delays of the gates of a sensor connected to a bit line exceeds a clock period, the sensor outputs an error, indicating an unsafe operation in the memory (or that normal memory operation is about to fail). In fact, the sensor detects a bit line transition and generates a pulse proportional to the transition time. Due to the presence of delay elements in the sensor, a delay is added to the pulse, which allows capturing the pulse delayed in a flip-flop and signaling the predictive error. If the delayed impulse is not captured by the flip-flop, there is no signaling.

As previously mentioned, this sensor is compatible with the use of DVFS techniques, as it can be used with low supply voltages. The purpose of using the sensor with this technique is to match each V_{DD} value with a different clock frequency, so that the sensor always works on the verge of an error, allowing the circuit to be operated at maximum performance, or minimum energy, ensuring that the safety margin remains similar for all V_{DD} levels.

This sensor can also be used internally in memory, as a local sensor (monitoring the real memory cells), or externally, as a global sensor, if it is placed to monitor a dummy memory cell.

Keywords: SRAM, DRAM, Performance sensor, memories, CMOS

Contents

List of Tables

List of Figures

Nomenclature

Greek symbols

λ	Device parameter.
μ_n	Mobility of the electrons at the surface of the n channel.
μ_p	Mobility of holes in the induced p channel.
ε_{ox}	Permittivity of silicon dioxide.

Roman symbols

A_v	Voltage gain.
C_{ox}	Oxide capacitance.
g_m	MOSFET transconductance.
i_D	Current drain.
$k_n^{'}$	Process transconductance parameter for NMOS.
$k_p^{'}$	Process transconductance parameter for PMOS.
k_n	Transconductance parameter for NMOS.
k_p	Transconductance parameter for PMOS.
L	Length of the MOSFET channel.
r_{DS}	Linear resistance of MOSFET in triode region.
r_o	Output resistance of MOSFET in saturation .
t_{ox}	Oxide thickness.
V_A	Early voltage.
v_{DS}	Voltage between drain and source.
v_{GD}	Voltage between gate and drain.
v_{GS}	Voltage between gate and source.

v_{ov} Overdrive voltage.

$V_{T,n}$ Threshold voltage for NMOS.

$V_{T,p}$ Threshold voltage for PMOS.

W_n Width of the channel for NMOS.

W_p Width of the channel for PMOS.

Glossary

BTI	Bias Temperature Instability
CLK	Clock
CMOS	Complementary Metal-Oxide Semiconductor
DRAM	Dynamic Random-Access Memory
DVFS	Dynamic Voltage and Frequency Scaling
DfT	Design for Testability
EM	Electromigration
HCI	Hot Carrier Injection
HW	Hardware
IC	Integrated Circuit
IoT	Internet of Things
MOSFET	Metal-Oxide Semiconductor Field-Effect Transistor
NBTI	Negative Bias Temperature Instability
NMOS	N-type Metal-Oxide Semiconductor
NMOSFET	N-type Metal-Oxide Semiconductor Field-Effect Transistor
PBTI	Positive Bias Temperature Instability
PMOS	P-type Metal-Oxide Semiconductor
PMOSFET	P-type Metal-Oxide Semiconductor Field-Effect Transistor
PMU	Power Management Unit
PVT	Process, power-supply Voltage and Temperature
PVTA	Process, power-supply Voltage ,Temperature and Aging
Si	Silicon (chemical symbol)
SoCs	Systems-on-a-Chip
SW	Software
SRAM	Static Random-Access Memory
TDDB	Time Dependent Dielectric Breakdown

Chapter 1

Introduction

Internet of Things (IoT) is stimulating the fourth industrial revolution, bringing significant benefits by connecting people, processes, and data [3][4]. The possibility of interconnecting a huge amount of smart hardware/software (hw/sw) systems, with increasing local artificial intelligence, is opening new avenues of research and innovative IoT applications across various markets, from smart cities [5] down to health systems [6], automotive applications [7], aerospace, and so on. On the opposite, IoT increases cybersecurity problems [8][9], causing reliability to be a key variable in new node hardware/software systems. What if some devices in an IoT net incorrectly store data? What if these elements (things) are driven to make “decisions” based on wrong data, or erroneous processing? And, due to the interconnectivity, what if such erroneous functionality triggers erroneous decisions in other IoT devices, thus spreading unsafe operation?

In the next section, hardware challenges in IoT are resumed, focusing performance and power related reliability issues.

1.1 Hardware Challenges in IoT

One crucial issue for IoT applications is power management [10][11]. A large array of smart sensors, often battery operated, requires high energy efficiency. This quests for ultra-low-power microcontrollers, and low-power memories. For this, a key variable is the minimum power supply voltage value, V_{DD}, which can guarantee safe data retention, and data access (read/write operations). Using a flexible Power Management Unit (PMU), it may be rewarding to perform Dynamic Voltage and Frequency Scaling (DVFS) to power memory arrays with minimum V_{DD} during memory access, and data retention, to reduce power consumption. And if more aggressive power savings are needed, techniques like power gating [12] can also be used to allow consumption only when is needed.

Systems-on-a-Chip (SoCs), and other integrated circuits, today are composed of nanoscale devices that are crammed in a very limited silicon area, presenting reliability issues and new challenges. CMOS circuits’ performance is sensitive to parametric variations, such as Process, power-supply Voltage and Temperature (PVT) [13], as well as aging effects (PVT and Aging – PVTA). CMOS circuit’s aging degradation is mainly caused by the following effects: Bias Temperature Instability (BTI), Hot-Carrier Injection (HCI), Electromigration (EM) and Time Dependent Dielectric Breakdown (TDDB) [14]. The most relevant aging effect is the BTI, namely the Negative Bias Temperature Instability (NBTI), which affects PMOS MOSFET transistors, resulting in a gradual increase of their absolute threshold voltages over time ($|V_{thP}|$). As high-k dielectrics started to be employed from the sub-32nm technologies [15], BTI also significantly affects NMOS transistors – Positive Bias Temperature Instability (PBTI), resulting in a rise of their threshold voltages, V_{thN}. These effects degrade digital circuit's performance over time, increasing the variability in CMOS circuits. Performance degradation decrease the switching speed, eroding time margins and leading to potential delay faults, and eventually chip failures.

Moreover, large nets of IoT devices are expensive, and are expected to operate for a long period of time. Hence, what if semiconductor aging phenomena cause an unacceptable device degradation, thus also causing incorrect functionality, during product lifetime?

This set of challenging issues pave the way to consider designing IoT devices with embedded monitoring capabilities, in a similar way that chip designers implement Design for Testability (DfT) techniques. Such non-mission functionality should allow to monitor local device operation, during product lifetime, and trigger warnings or corrective decisions, in order to guarantee safe operation, in a safety level adequate to the IoT application. It can also be used to perform DVFS on microcontrollers and memory banks, thus allowing considerable power savings.

Moreover, today's SoC face the rapidly increasing need to store more and more information. As a consequence, memories occupy the greatest part of the SoC silicon area, being currently around 90% of SoC density [16]. Therefore, memories' robustness is considered crucial in order to guarantee the reliability of such SoCs over product lifetime [16]. The trend is that this predominance of memory *Si* area on logic *Si* area will continue to grow in the following years. Consequently, semiconductor memory has become the main responsible of the overall SoC area, and for the active and leakage power in embedded systems and, thus, in the hardware part of IoT devices.

One of the major issues in the design of an SRAM cell is stability. Cell stability is basically its ability to maintain correct operation in the presence of noise signals, thus ensuring correct Read, Write and Hold operations. Therefore, it determines the sensitivity of the memory to parametric variations, induced in the manufacturing process and/or during operating conditions. Static noise margins (SNMs) are widely used as the criteria of stability [17]. However, some authors defend that dynamic noise margins are also important [18]. Nevertheless, due to PVTA variations (and knowing that aging is a cumulative process), a degradation in memory's performance and stability may occur.

1.2 Motivation

In the past, significant research has been carried out, and a set of cost-effective performance sensors for digital logic, either in a cell-library design style in custom SoCs, or in an FPGA programmable tissue has been proposed (see, e.g., [19][20][21]). However, research on performance sensors for semiconductor memories has been much more limited, so far.

We acknowledge that there is vast previous research dealing with aging sensors for SRAM cells, and especially focused on the BTI effect. These are attempts to increase reliability in SRAM operation. Nevertheless, they do not simultaneously consider PVT and Aging variations. Hence, previous work mainly deals with sensing aging in SRAMs, but a cost-effective generic sensor to deal with performance and, simultaneously, PVTA variations in memories is missing. Moreover, previous work does not address the development of SRAM sensors for ultra-low-power operation, a mandatory request for many IoT applications. Regarding previous works on DRAM, the available works are even more limited when compared with SRAM related works.

The common problem is that all possible circuit variations (namely, PVTA variations) affect cumulatively circuit performance and behaviour, and due to their cumulative effect, may be responsible for error occurrence, thus compromising safe IoT operation. Therefore, it is important to develop a sensor that can be aware to all these time variations, i.e., a performance sensor. In fact, research on performance sensors for digital synchronous logic is much more ahead when compared with their memory counterparts. As an example, the Scout Flip-Flop sensor [22][19] acts as a performance sensor for tolerance and predictive detection of delay faults in synchronous digital circuits (ASIC).

However, research on on-line SRAM sensors which may identify abnormal time response, regardless of its origin, is still limited. In fact, as far as the authors knowledge, there are only three previous works on performance sensors ([23], [24], and [2][25][26]), which are an initial attempt to develop a performance sensor for memories. Unfortunately, the sensor architecture, proposed in [23] is complex and leads to a significant area overhead, and the sensor's performance is limited in the presence of reduced V_{DD} voltages. The sensor proposed in [24], although resolves part of the problems referred, but it still has implementation issues that prohibits its use in a real memory circuit. Finally, the work proposed in [2][25][26], although it presents sensor versions for both SRAM and DRAM memories, and it is a performance sensor (sensitive to PVTA variations), it fails on working in subthreshold voltages, which is a key aspect for new node IoT chips.

1.3 Objectives

The first objective of this work was to study the applicability of work described in [2][25][26], the Scout Memory Sensor, to ultra-low-power circuits, i.e., to study its use under subthreshold power-supply voltage values. This study should allow to define the limits of minimum V_{DD} voltage that could be used with correct operation of the sensor. Regardless of the results that would be obtained with this study, it was already known for a fact that the minimum working V_{DD} is not a subthreshold voltage level and that this sensor should be improved.

The second objective, and the main purpose of this thesis, was to propose changes to the Scout Memory Sensor that could define a novel, ultra-low power, and on-line performance sensor for SRAM and/or DRAM circuits, targeting IoT applications. The PVTA-aware performance sensor should allow to detect timing degradation on the access to CMOS memory cells, namely in read/write operations, and it should do it even at subthreshold power supply voltage values. The sensor should be connected to the memories' bit lines, to monitor transitions occurred in these signals during these read/write operations. It should on-line monitor any performance variation with a very low performance overhead and a reasonable area overhead. The aging and/or performance monitoring should be achieved by detecting slow transitions due to a reduction of performance caused by PVTA variations (or by any other time response degrading effect) in the memory cells or in the memory circuitry (like in the sense amplifier, also connected to the bit lines).

Furthermore, a final objective of this work was that the time response degradation of a memory circuit with the sensor could be carried out on purpose, to constraint power consumption. Hence, it should be analysed the possibility to use the sensor to tune a Dynamic Voltage and Frequency Scaling methodology, by signalling to the PMU the lowest V_{DD} value which can be used to correctly perform memory access, within a user's defined time safety margin. Of course, to guarantee these two last objectives, the sensor's architecture must be designed to guarantee the sensor's correct operation under these ultra-low V_{DD} values.

1.4 Context of The Research Work

The research and development of this master's thesis was carried out at the Instituto Superior Técnico (IST) of the University of Lisbon (UL) in collaboration with INESC-ID in Lisbon and University of Algarve – Engineering Institute in Faro.

This work is part of the Master's program in Aerospace Engineering with specialization in the field of avionics and Minor in electronics and telecommunications.

1.5 Thesis Outline

This thesis is organized in the following chapters:

Chapter 2: **CMOS Memories and PVTA Variations** presents the structure and architecture CMOS memories, in particular SRAM and DRAM memories. It also has a brief introduction to PVTA variations, focusing mainly on the NBTI and PBTI aging effects and their influence on the performance of memories. Process, power-supply voltage and temperature variations are described. Finally, an analysis of subthreshold techniques is carried out in order to justify their use in sensors, in particular the importance of energy saving and the existence of a compromise between power and performance.

Chapter 3: **State of the Art on Performance Sensors** presents some background works on aging and performance sensors illustrating their architectures and main characteristics. In this chapter is described the OCAS sensor, that consists of an aging sensor sensitive to NBTI effects. A performance sensor for SRAM memories performed by Hugo Santos [23][24] and an improved version of the same sensor (the Scout memory sensor) with application in SRAM and DRAM memories performed by Luis Santos [2][25][26]. Finally, a sensor for synchronous logic circuits is described, which consists of two types of sensors, local performance sensor (LPS) and global performance sensor (GPS) and which uses an Adaptive Voltage Scaling (AVS) strategy to ensure its applicability under subthreshold conditions.

Chapter 4: **Study of Scout Memory Sensor for subthreshold voltages** aims to analyze the operation of the Scout Memory Sensor for supply voltages below the nominal voltage, as is the case of the subthreshold regime. This chapter aims to know if for lower voltages, the Scout Memory Sensor is still a robust solution, presenting a reliable behavior and which minimum supply voltage the sensor is reliable. In this chapter some parametric simulations are presented for each block of the Scout Memory Sensor, namely the transition detector, the pulse detector and the complete circuit of the sensor. Finally, a brief conclusion is made about the scout memory sensor's accuracy for lower supply voltages, mentioning some advantages and disadvantages of this sensor.

Chapter 5: **Ultra-Low Power Performance Sensor for Memories** aims to present a new sensor that allows to overcome the problems detected in the Scout Memory Sensor, when supply voltages below the nominal voltage are used. In this chapter is presented the architecture of this new sensor consisting of three blocks (transition detector, delay element, and flip-flop). In each block simulations are carried out in Cadence allowing to observe the operation of each sensor structure. The final sensor circuit and some simulations are then displayed. Finally, a usability analysis of the sensor is made, focusing mainly on its use as a local sensor and global sensor, along with DVFS techniques. Some advantages and disadvantages of this type of sensor are also described.

Chapter 6: **Layouts and Simulation Results** presents the layouts developed for all the blocks of the new sensor implementations, as well as test circuits. The simulations performed under Cadence framework for all the circuits are also presented here, and the main results are analyzed.

Chapter 7: **Conclusions** resumes the main conclusions and summarizes the future work.

Chapter 2

CMOS Memories and PVTA Variations

In this chapter we will first make a brief introduction to CMOS memories, focusing on the structure of SRAM and DRAM memories, because this work refers to sensors applicable to this type of memories, and we will also describe peripheral circuits and sense amplifiers of memories. PVTA effects affecting the performance of CMOS circuits are then summarized. The last section of this chapter describes the use of subthreshold techniques that allow sensors to save more energy, making the process more efficient.

2.1 CMOS Memories

Semiconductor memory arrays capable of storing large quantities of digital information are essential to all digital systems. The amount of memory required in a particular system depends on the type of application, but, in general, the number of transistors used for the information (data) storage function is much larger than the number of transistors used in logic operations and for other purposes. The ever-increasing demand for larger data storage capacity has driven the fabrication technology and memory development towards more compact design rules and, consequently, toward higher data storage densities. Thus, the maximum realizable data storage capacity of single-chip semiconductor memory arrays approximately doubles every two years. On-chip memory arrays have become widely used subsystems in many VLSI circuits, and commercially available single-chip read/write memory capacity has reached 64 megabits. This trend toward higher memory density and larger storage capacity will continue to push the leading edge of digital systems' design.

The area efficiency of the memory array, i.e., the number of stored data bits per unit area, is one of the key design criteria that determine the overall storage capacity and, hence, the memory *cost per bit*. Another important issue is the memory access time, i.e., the time required to store and/or retrieve a particular data bit in the memory array. The access time determines the memory speed, which is an important performance criterion of the memory array. Finally, the static and dynamic power consumption of the memory array is a significant factor to be considered in the design, because of the increasing importance of low-power applications.

Memory circuits are generally classified according to the type of data storage and the type of data access. *Read-Only Memory* (ROM) circuits allow, as the name implies, only the retrieval of previously stored data and do not permit modifications of the stored information contents during normal operation. ROMs are *non-volatile* memories, i.e., the data storage function is not lost even when the power supply-voltage is off.

Read-write (R/W) memory circuits, on the other hand, must permit the modification (writing) of data bits stored in the memory array, as well as their retrieval (reading) on demand. This requires that the data storage function be *volatile*, i.e., the stored data are lost when the power supply voltage is turned off. The read-write memory circuit is commonly called *Random Access Memory* (RAM), mostly due to historical reasons. Compared to sequential-access memories such as magnetic tapes, any cell in the R/W memory array can be accessed with nearly equal access time. Based on the operation type of individual data storage cells, RAMs are classified into two main categories: *Static* RAMs (SRAM) and *Dynamic* RAMs (DRAM).

A typical memory array organization is shown in figure 2.1. The data storage structure, or core, consists of individual memory cells arranged in an array of horizontal rows and vertical columns. Each *cell* is capable of storing one bit of binary information. Also, each memory cell shares a common connection with the other cells in the same row, and another common connection with the other cells in the same column. In this structure, there are 2^N rows, also called *word lines*, and 2^M columns, also called *bit lines*. Thus, the total number of memory cells in this array is $2^M \times 2^N$.

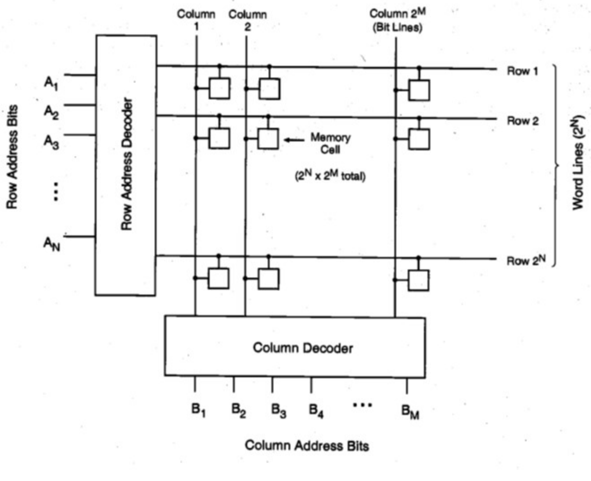


Figure 2.1: Typical RAM array organization.

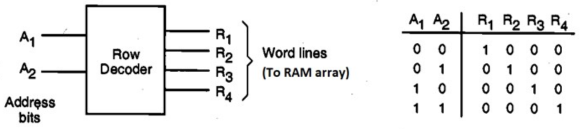
To access a particular memory cell, i.e., a particular data bit in this array, the corresponding bit line and the corresponding word line must be activated (selected). The row and column selection operations are accomplished by row and column *decoders*, respectively. The row decoder circuit selects one out of 2^N word lines according to an N-bit row address, while the column decoder circuit selects one out of 2^M bit lines according to an M-bit column address. Once a memory cell or a group of memory cells are selected in this fashion, a data read and/or a data write operation may be performed on the selected single bit or multiple bits on a particular row. The column decoder circuit serves the double duties of selecting the particular columns and routing the corresponding data content in a selected row to the output.

We can see from this simple discussion that individual memory cells can be accessed for data read and/or data write operations in random order, independent of their physical locations in the memory array. Thus, the array organization examined here is called a *Random Access Memory* (RAM) structure.

2.1.1 Peripheral Circuits

Row-Address Decoder

Now we will turn our attention to the circuit structures of row and column address decoders, which select a particular memory location in the array, based on the binary row and column addresses. A row decoder designed to drive a NOR RAM array must, by definition, select one of the 2^N word lines by raising its voltage to V_{OH}. As an example, consider the simple row address decoder shown in figure 2.2, which decodes a two-bit row address and selects one out of four word lines by raising its level.



A_1	A_2	R_1	R_2	R_3	R_4
0	0	1	0	0	0
0	1	0	1	0	0
1	0	0	0	1	0
1	1	0	0	0	1

Figure 2.2: Row address decoder example for 2 address bits and 4 word lines.

A most straightforward implementation of this decoder is another NOR array, consisting of 4 rows (outputs) and 4 columns (two address bits and their complements). Note that this NOR-based decoder array can be built just like the NOR RAM array, using the same selective programming approach (figure 2.3).

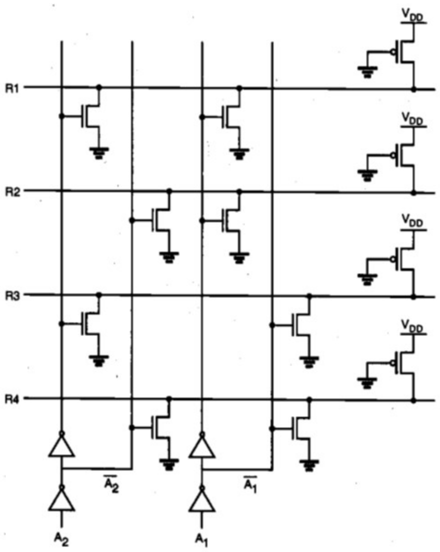


Figure 2.3: NOR-based row decoder circuit for 2 address bits and 4 word lines.

Column-Address Decoder

The column decoder circuitry is designed to select one out of 2^M bit lines (columns) of the RAM array according to an M-bit column address, and to route the data content of the selected bit line to the data output. A straightforward but costly approach would be to connect an NMOS pass transistor to each bit-line (column) output, and to selectively drive one out of 2^M pass transistors by using a NOR-based column address decoder, as shown in figure 2.4. In this arrangement, only one NMOS pass transistor is turned on at a time, depending on the column address bits applied to the decoder inputs. The conducting pass transistor routes the selected column signal to the data output. Similarly, a number of columns can be chosen at a time, and the selected columns can be routed to a parallel data output port.

Note that the number of transistors required for this column decoder implementation is $2^M(M+1)$, i.e., 2^M pass transistors for each bit line and $M2^M$ transistors for the decoder circuit. This number can quickly become excessive for large M, i.e., for a large number of bit lines.

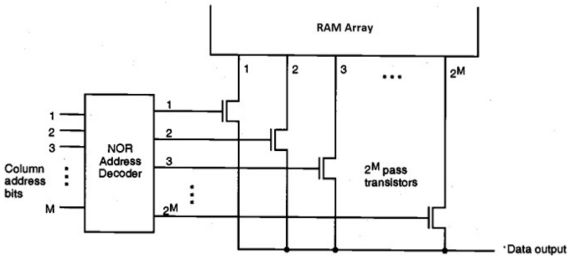


Figure 2.4: Bit-line (column) decoder arrangement using a NOR address decoder and NMOS pass transistors for every bit line.

An alternative design of the column decoder circuit is to build a binary selection tree consisting of consecutive stages, as shown in figure 2.5. In this case, the pass transistor network is used to select one out of every two bit lines at each stage (level), whereas the column address bits drive the gates of the NMOS pass transistors. Notice that a NOR address decoder is not needed for this decoder tree structure, thereby reducing the number of transistors significantly although it requires *M* additional inverters (*2M* transistors) for complementing column address bits. The example shown in figure 2.5 is a column decoder tree for eight bit lines, which requires three column address bits (and their complements) to select one of the eight columns.

One drawback of the decoder tree approach is that the number of series-connected NMOS pass transistors in the data path is equal to the number of column address bits, *M*. This situation can cause a long data access time, since the decoder delay time depends on the equivalent series resistance of the decoder branch that directs the column data to the output.

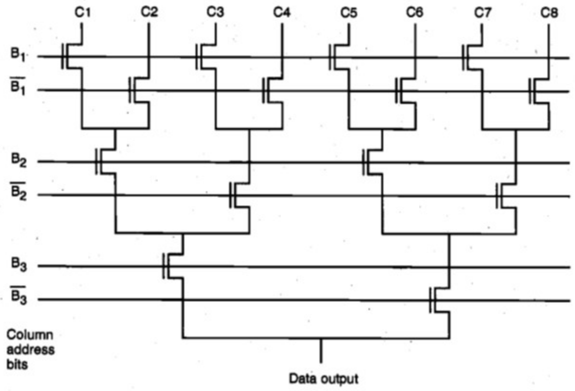


Figure 2.5: Column decoder circuit for eight bit lines, implemented as a binary tree decoder which is driven directly by the three column address bits.

Precharge and Equalization

The *precharging* of bit lines also plays a significant role in the access time. In an unclocked RAM array, data from the accessed cell develops a voltage difference on the bit lines. This voltage difference is then detected and amplified to drive the output buffer. When another cell on the same column is accessed next, one that contains data opposite to the data contained in the previously accessed cell, the output has to switch first to an equalized state and then to the opposite logic state. Since the capacitance on the bit lines is quite large, the time required for switching the differential from one state to the other becomes a significant portion of the overall access time. The access time penalty associated with this procedure can be substantially reduced by the *equalization* of bit lines prior to each new access. Equalization can be done when the memory array is deselected, i.e., between two access cycles.

2.1.2 Sense Amplifiers

Figure 2.6 shows the sense amplifier together with some of the other column circuitry of a RAM chip. Note that the sense amplifier is nothing but the familiar latch formed by cross-coupling two CMOS inverters: One inverter is implemented by transistors Q_1 and Q_2, and the other by transistors Q_3 and Q_4. Transistors Q_5 and Q_6 act as switches that connect the sense amplifier to ground and V_{DD} only when data-sensing action is required. Otherwise, φ_S is low and the sense amplifier is turned off. This conserves power, an important consideration because usually there is one sense amplifier per column, resulting in thousands of sense amplifiers per chip. Note, again, that terminals *x* and *y* are both the input and the output terminals of the amplifier. As indicated, these I/O terminals are connected to the B and $\overline{B}$ lines. The amplifier is required to detect a small signal appearing between B and $\overline{B}$, and to amplify it to provide a full-swing signal at B and $\overline{B}$. For instance, if during a read operation, the cell has a stored 1, then a small positive voltage will develop between B and $\overline{B}$, with v_B higher than $v_{\overline{B}}$. The amplifier will then cause v_B to rise to V_{DD} and $v_{\overline{B}}$ to fall to 0 V. This 1 output is then directed to the chip I/O pin by the column decoder and at the same time is used to rewrite a 1 in the DRAM cell, thus performing the restore operation that is required because the DRAM readout process is destructive.

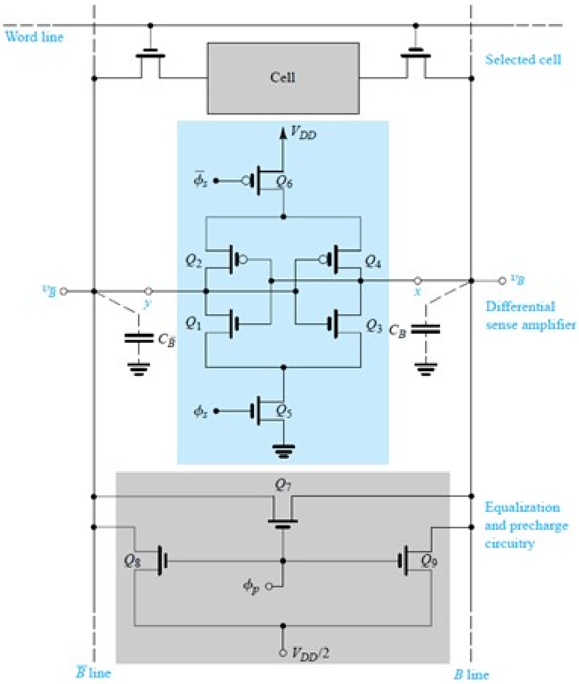


Figure 2.6: A differential sense amplifier connected to the bit lines of a particular column.

2.1.3 SRAM

Read-write (R/W) memory circuits are designed to permit the modification (writing) of data bits to be stored in the memory array, as well as their retrieval (reading) on demand. The memory circuit is said to be static if the stored data can be retained indefinitely (as long as a sufficient power supply voltage is provided), without any need for a periodic refresh operation.

Full CMOS SRAM Cell

A low-power SRAM cell may be designed simply by using cross-coupled CMOS inverters. In this case, the stand-by power consumption of the memory cell will be limited to the relatively small leakage currents of both CMOS inverters. The possible drawback of using CMOS SRAM cells, on the other hand, is that the cell area tends to increase in order to accommodate the n-well for the PMOS transistors and the polysilicon contacts.

The circuit structure of the full CMOS static RAM cell is shown in figure 2.7, along with the PMOS column pull-up transistors on the complementary bit lines. The most important advantage of this circuit topology is that the static power dissipation is even smaller; essentially, it is limited by the leakage current of the PMOS transistors. A CMOS memory cell thus draws current from the power supply only during a switching transition. The low standby power consumption has certainly been a driving force for the increasing prominence of high- density CMOS SRAMs.

Other advantages of CMOS SRAM cells include high noise immunity due to larger noise margins, and the ability to operate at lower power supply voltages than, for example, the resistive-load SRAM cells. The major disadvantages of CMOS memories historically were larger cell size, the added complexity of the CMOS process, and the tendency to exhibit ”latch-up” phenomena. With the widespread use of multi-layer polysilicon and multi-layer metal processes, however, the area disadvantage of the CMOS SRAM cell has been reduced significantly in recent years. Considering the undisputable advantages of CMOS for low-power and low-voltage operation, the added process complexity and the required latch-up prevention measures do not present a substantial barrier against the implementation of CMOS cells in high density SRAM arrays.

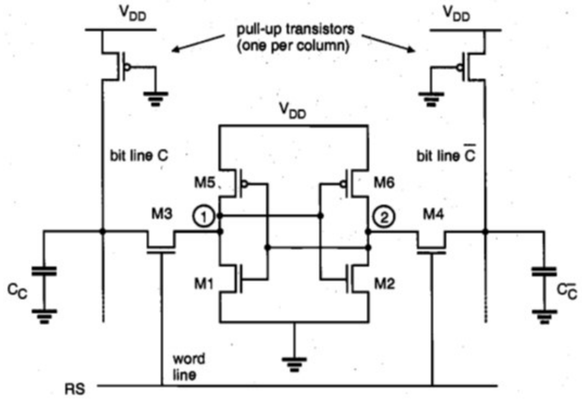


Figure 2.7: Circuit topology of the CMOS SRAM cell.

Read Operation

Consider the data-read operation first, assuming that a logic ”0” is stored in the cell. The voltage levels in the CMOS SRAM cell at the beginning of the ”read” operation are depicted in figure 2.8. Here, the transistors M2 and M5 are turned off, while the transistors M1 and M6 operate in the linear mode. Thus, the internal node voltages are $V_1 = 0$ and $V_2 = V_{DD}$ before the cell access (or pass) transistors M3 and M4 are turned on. The active transistors at the beginning of the data-read operation are highlighted in figure 2.8.

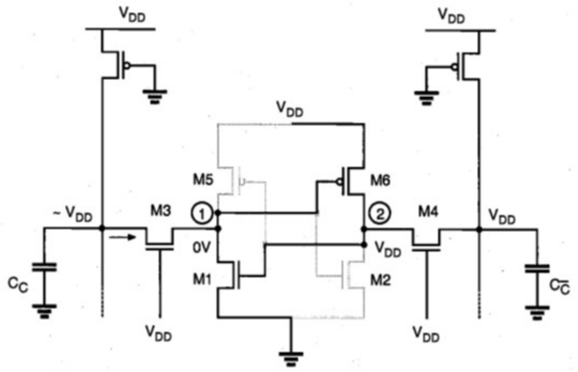


Figure 2.8: Voltage levels in the SRAM cell at the beginning of the ”read” operation.

After the pass transistors M3 and M4 are turned on by the row selection circuitry, the voltage level of column $\overline{C}$ will not show any significant variation since no current will flow through M4. On the other half of the cell, however, M3 and M1 will conduct a nonzero current and the voltage level of column C will begin to drop slightly. Note that the column capacitance C_C is typically very large; therefore, the amount of decrease in the column voltage is limited to a few hundred millivolts during the read phase. The data-read circuitry is responsible for detecting this small voltage drop and amplifying it as a stored ”0”. While M1 and M3 are slowly discharging the column capacitance, the node voltage V_1, will increase from its initial value of 0 V. Especially if the (W/L) ratio of the access transistor M3 is large compared to the (W/L) ratio of M1, the node voltage V_1 may exceed the threshold voltage of M2 during this process, forcing an unintended change of the stored state. The key design issue for the data-read operation is then to guarantee that the voltage V_1, does not exceed the threshold voltage of M2, so that the transistor M2 remains turned off during the read phase, i.e.,

$$V_{1,max} \leq V_{T,2} \tag{2.1}$$

We can assume that after the access transistors are turned on, the column voltage V_C remains approximately equal to V_{DD}. Hence, M3 operates in saturation while M1 operates in the linear region.

$$\frac{k_{n,3}}{2}\left(V_{DD} - V_1 - V_{T,n}\right)^2 = \frac{k_{n,1}}{2}\left(2\left(V_{DD} - V_{T,n}\right)V_1 - V_1^2\right) \tag{2.2}$$

Combining this equation with (2.1) results in:

$$\frac{k_{n,3}}{k_{n,1}} = \frac{\left(\frac{W}{L}\right)_3}{\left(\frac{W}{L}\right)_1} < \frac{2\left(V_{DD} - 1,5V_{T,n}\right)V_{T,n}}{\left(V_{DD} - 2V_{T,n}\right)^2} \tag{2.3}$$

Write Operation

Now consider the write ”0” operation, assuming that a logic ”1” is stored in the SRAM cell initially. Figure 2.9 shows the voltage levels in the CMOS SRAM cell at the beginning of the data-write operation. The transistors M1 and M6 are turned off, while the transistors M2 and M5 operate in the linear mode. Thus, the internal node voltages are $V_1 = V_{DD}$ and $V_2 = 0$ V before the cell access (or pass) transistors M3 and M4 are turned on.

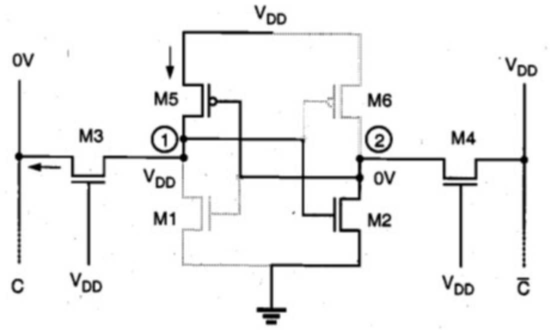


Figure 2.9: Voltage levels in the SRAM cell at the beginning of the ”write” operation.

The column voltage V_C is forced to logic ”0” level by the data-write circuitry; thus, we may assume that V_C is approximately equal to 0 V. Once the pass transistors M3 and M4 are turned on by the row selection circuitry, we expect that the node voltage V_2 remains below the threshold voltage of M1, since M2 and M4 are designed according to condition (2.3). Consequently, the voltage level at node (2) would not be sufficient to turn on M1. To change the stored information, i.e., to force V_1 to 0 V and V_2 to V_{DD}, the node voltage V_1 must be reduced below the threshold voltage of M2, so that M2 turns off first. When $V_1 = V_{T,n}$, the transistor M3 operates in the linear region while M5 operates in saturation.

$$\frac{k_{p,5}}{2}\left(0 - V_{DD} - V_{T,p}\right)^2 = \frac{k_{n,3}}{2}\left(2\left(V_{DD} - V_{T,n}\right)V_{T,n} - V_{T,n}^2\right) \tag{2.4}$$

Rearranging this condition results in:

$$\frac{k_{p,5}}{k_{n,3}} < \frac{2\left(V_{DD} - 1,5V_{T,n}\right)V_{T,n}}{\left(V_{DD} + V_{T,p}\right)^2} \Leftrightarrow \frac{\left(\frac{W}{L}\right)_5}{\left(\frac{W}{L}\right)_3} < \frac{\mu_n}{\mu_p}\frac{2\left(V_{DD} - 1,5V_{T,n}\right)V_{T,n}}{\left(V_{DD} + V_{T,p}\right)^2} \tag{2.5}$$

2.1.4 DRAM

As the trend for high-density RAM arrays forces the memory cell size to shrink, alternative data storage concepts must be considered to accommodate these demands. In a dynamic RAM cell, binary data is stored simply as charge in a capacitor, where the presence or absence of stored charge determines the value of the stored bit. Note that the data stored as charge in a capacitor cannot be retained indefinitely, because the leakage currents eventually remove or modify the stored charge. Thus, all dynamic memory cells require a periodic refreshing of the stored data, so that unwanted modifications due to leakage are prevented before they occur. The use of a capacitor as the primary storage device generally enables the DRAM to be realized on a much smaller silicon area compared to the typical SRAM cell.

One-Transistor DRAM Cell

The circuit diagram of the one-transistor (1-T) DRAM cell consisting of one explicit storage capacitor and one access transistor is shown in figure 2.10. Here, C_1 represents, the storage capacitor which typically has a value of 30 fF to 100 fF and binary data are stored as the presence or absence of charge in the storage capacitor. Capacitor C_2 represents the much larger parasitic column capacitance associated with the word line. Charge sharing between this large capacitance and the very small storage capacitance plays a very important role in the operation of the 1-T DRAM cell.

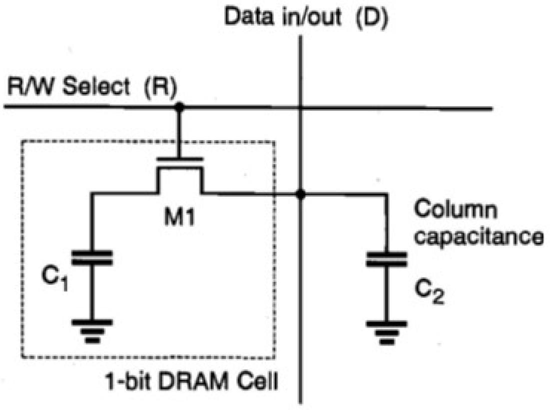


Figure 2.10: Typical one-transistor (1-T) DRAM cell with its access lines.

Write Operation

The ”data write” operation on the 1-T cell is quite straightforward. For the write ”1” operation, the bit line (D) is raised to logic ”1” by the write circuitry, while the selected word line is pulled high by the row address decoder. The access transistor M1 turns on, allowing the storage capacitor C_1 to charge up to a logic-high level. For the write ”0” operation, the bit line (D) is pulled to logic ”0” and the word line is pulled high by the row address decoder. In this case, the storage capacitor C_1 discharges through the access transistor, resulting in a stored ”0” bit.

Read Operation

In order to read stored data out of a 1-T DRAM cell, on the other hand, we have to build a fairly elaborate read-refresh circuit. The reason for this is the fact that the ”data read” operation on the one-transistor DRAM cell is by necessity a ”destructive readout”. This means that the stored data must be destroyed or lost during the read operation. Typically, the read operation starts with precharging the column capacitance C_2. Then, the word line is pulled high in order to activate the access transistor M1. Charge sharing between C_1 and C_2 occurs and, depending on the amount of stored charge on C_1, the column voltage either increases or decreases slightly. Note that charge sharing inevitably destroys the stored charge on C_1. Hence, we also have to refresh data every time we perform a ”data read” operation.

2.2 PVTA Variations

2.2.1 Aging Variation

The challenges of designing integrated circuits (ICs) are focused on accomplishing high reliability and performance, which are partially associated with minimizing aging effects. In MOS technology, the degradation phenomena are classified as destructive and non-destructive, with bias-temperature instability (BTI) and hot-carrier injection (HCI) being non-destructive cases and manifesting themselves as charge carrier tunneling from the inversion channel into the gate's dielectric, due to the continuous increasing of vertical and horizontal electric fields. Destructive degradation manifests as electromigration (EM) and time-dependent dielectric breakdown (TDDB), destroying the physical and electrical functionality of interconnections and the MOS's gate insulator [22].

This chapter reviews the most important integrated-circuit aging phenomena's, in special the bias temperature instability (BTI) effects: negative bias temperature instability (NBTI) and positive bias temperature instability (PBTI).

BTI Effect

BTI is commonly associated with an increase of MOSFET devices' threshold voltage (V_T), which leads to charge carrier's mobility reduction within the conduction channel, ultimately reducing drain current and the transistor's transconductance. Even though this phenomenon has been known for almost fifty years, its complete understanding remains a mystery. Nevertheless, BTI degradation mechanisms can be associated with interface traps' (N_{it}) generation, also known as the P_b center, oxide charge (N_{ot}), and pre-existent defects within the dielectric layer, or oxygen vacancies (O_v) [16] occupancy due to charge tunneling from the inversion channel. Due to the continuous increase of V_T, each time, a higher gate-voltage is needed to obtain the prior-to-stress overdrive voltage. In old technologies, e.g., the length (L) of the channel wider than 90 nm, BTI was only considered on p-type MOS (PMOS) transistors because its impact on n-type MOS (NMOS) devices is almost negligible. BTI is known as negative (NBTI) for the case of PMOS transistors and positive (PBTI) for NMOS transistors [20].

BTI's leading mechanism is trap generation at the interface between the substrate and silicon dioxide, so the natural question is: What are these interface traps, and how can they affect the functionality of MOSFETs?

When silicon oxidizes, the bonding configuration at the surface will depend on the wafer's crystallographic orientation; while most of the silicon atoms bond to an oxygen atom, some others might bond to hydrogen atoms (the element used for the passivation of point defects during the manufacturing process). An interface trap, also known as the P_b center, consists of a silicon atom at Si/SiO_2 interface that has only three complete bonds and an unsatisfied fourth bond, known as a dangling bond. That unoccupied bond is perpendicular to the interface and points towards an oxygen vacancy located above itself [20]. These interface traps are the result of the mismatch between Si/SiO_2 at the interface due to the generated stress during the gate insulator's thermal growth; these interface traps are capable of trapping charge carriers from the conduction channel. Interface traps are electrically-charged defects throughout the band-gap of silicon acting as generation/recombination centers contributing to leakage currents' increase and, further, charge carriers' mobility reduction. While in the upper half of silicon's band-gap, interface traps are the acceptor-like type, below the mid-gap, they behave as donor-like. Acceptor-like defects placed above the intrinsic Fermi level (E_i) and below the conduction-band level (E_C) have a neutral charge, as well as donor-like defects below the intrinsic level and above the valence-band level (E_V), for the case of intrinsic semiconductor materials.

On PMOS, where the bulk is doped with donor-type impurities, the Fermi level gets closer to the conduction band, so those energy-states between the intrinsic and Fermi levels become negatively charged. When the gate-voltage of a MOSFET device is strong enough to surpass the flat-band condition, the bulk's energy-bands will slightly bend upwards. When the strong inversion condition is achieved, the intrinsic level will bend below the Fermi level, and the originally negatively-charged states will become neutral again. By the time the device is taken to deep inversion, the intrinsic level will bend above E_F, so states between E_i and E_F will become positively charged, meaning that bonds at the interface have broken. For the case of NMOS devices, in which the substrate is acceptor-type doped, the interface trap's generation is the inverse process, considering that the bias condition for an NMOS is positive at the gate electrode, so the energy bands will bend downwards, and those states placed between Fermi and intrinsic level will become negatively charged, trapping electrons instead of holes [20].

A second BTI mechanism is oxide charge defects, which are charged impurities deposited into the gate oxide during the manufacturing process, K^+ and/or Na^+ ions. Nitrogen atoms are commonly used to passivate defects within the gate insulator; one drawback is that nitrogen creates nitrogen-rich layers within the gate insulator, becoming temporary charge traps during stress [20]. The last subtype of oxide charges comprises those stress-generated defects in the gate's dielectric. All oxide charges have less impact than interface traps. However, depending on oxide charges' location, the vertical electric field can be modified, further increasing the threshold voltage shift.

BTI is commonly modeled by using the previously-explained mechanisms for silicon technology, comprehended as the n- or p-doped silicon substrate, silicon dioxide as the gate insulator, and polysilicon as the gate electrode. As shown in Figure 2.11 [20], PBTI is ignored on micro-metric Technologies due to its minimal impact in NMOS devices, if compared to PMOS NBTI. Besides, in modern nanometer technologies based on high-*k* gate oxides and metal alloys as the gate electrode, PBTI takes a new degradation level, surpassing the overall threshold voltage degradation in PMOS transistors.

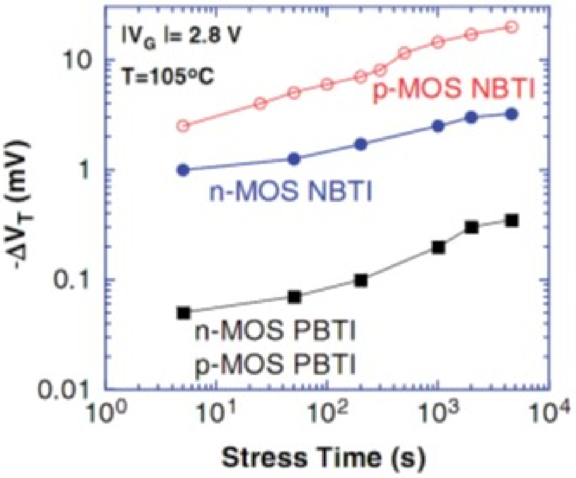


Figure 2.11: Comparison of NBTI and PBTI in both NMOS and PMOS devices.

High-*k* and metal gate technology (HKMG) has an advantage over the silicon-based one, keeping the scaling down tendency of the electric inversion layer (T_{inv}), maintaining short-channel effects at the check, and reducing leakage currents. However, as in silicon-based technology, the complete understanding of BTI degradation mechanisms, as well as their respective locations within the gate stack remains unknown. Nevertheless, researchers agree that the possible mechanisms behind BTI's appearance are: (1) the generation of interface states between the interfacial layer and substrate ($N_{it,il}$) and between the interfacial layer and high-*k* gate dielectric ($N_{it,hk}$), (2) the continuous trapping and detrapping of holes/electrons into preexisting oxygen vacancies on either the high-*k* layer (HK) or interface layer (IL), and (3) defect generation within the dielectric layer during stress (N_{ot}).

By 1977, the reaction-diffusion (RD) model [19] was more than enough to fit the measured BTI data with an analytical equation in a simulator and to be able to predict the BTI behavior, even though the discovery of the recovery phase showed the need for model changing due to the incapability of the RD model to predict BTI correctly under AC stress. In a physics-based BTI model, the mechanisms taken into account are interface traps' generation, charge carriers trapping on either preexistent or stress-generated defects within the oxide layers, and oxide charges. For instance, the threshold voltage change for the RD model is described by equation (2.6), which includes stress time and operation temperature, N_{ox} being the positive charge density within the oxide layer, N_{it} the initial interface traps' density, K_{ox} the relative constant of the gate dielectric, t_{ox} the oxide thickness, and ϵ_0 the vacuum's permittivity.

$$\Delta V_{T,DC} = -\frac{\Delta Q_{ox} + \Delta Q_{it}}{C_{ox}} = -\frac{q\left(\Delta N_{ox} + \Delta N_{it}\right)}{K_{ox}\epsilon_0} t_{ox} \tag{2.6}$$

NBTI

In [17], NBTI degradation was modeled by using three uncorrelated components: (1) interface trap generation at both the substrate and interfacial layer (Si/IL) and the interfacial and high-*k* dielectric layer (IL/HK) interfaces, (2) hole trapping into preexistent defects within the interfacial layer, and (3) the generation of new traps within the interfacial layer due to stress. The interface trap component is modeled using the reaction-diffusion model [19], while hole trapping uses the two energy well (2EW) or multi-state model [16].

Figure 2.12 shows the time evolution for NBTI's interface traps' generation in accordance with [10][17]. It is described as follows: the holes' inversion layer at the MOSFET's conduction channel breaks Si-H bonds located at the Si/IL interface; the released hydrogen atoms diffuse from the interfacial transition layer (SiO_x) and within IL; once at the Si/IL interface, these hydrogen atoms react with more hydrogen atoms breaking O_v-H (oxygen vacancies passivated with hydrogen) bonds. As in the RD model, dangling bonds are created, and H_2 molecules form within the interfacial layer, so at longer stress times, NBTI dynamics will be ruled by the diffusion of H_2 molecules diffusing from the Si/IL to IL/HK interface.

In [18], the differential equations describing interface traps' generation are presented, as well as the simplified form of those equations is found in [17]. Equation (2.7) describes the threshold voltage change in nano-metric MOS devices, C_{ox} being the device's gate capacitance, ΔN_{IT-IL} the generation of interface traps, ΔN_{HT-IL} the hole trapping related to the degradation process, and ΔN_{OT-IL} the bulk trap generation at the Si/IL interface.

$$\Delta V_T = \frac{q}{C_{ox}} \left(\Delta N_{IT-IL} + \Delta N_{HT-IL} + \Delta N_{OT-IL} \right) \tag{2.7}$$

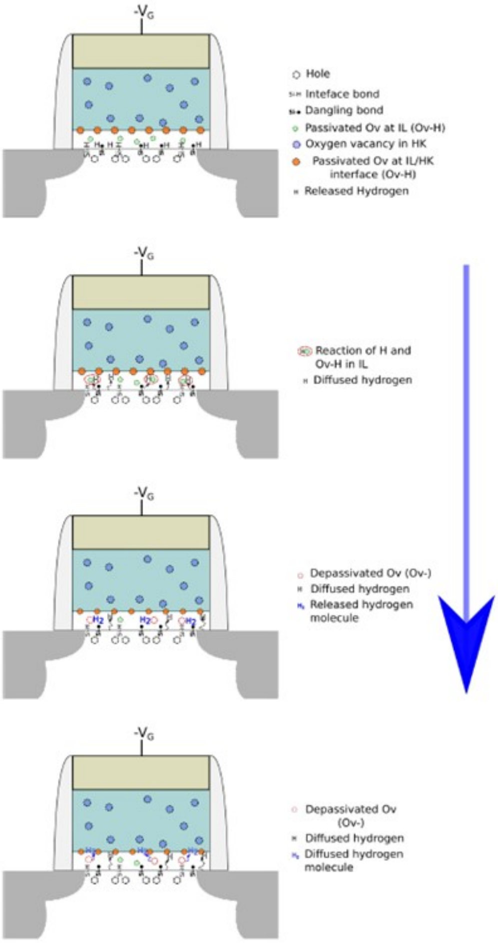


Figure 2.12: Degradation time evolution for NBTI interface traps' generation.

PBTI

The simplified model for PBTI degradation on NMOS transistors was provided in [10], including an intensive study of several devices under different manufacturing conditions. Stress-induced leakage currents on nano-metric NMOS transistors have been attributed to oxygen vacancies within the HK dielectric [4][5], which become charge traps for those charge carriers whose energy is close to the conduction-band energy of the high-*k* dielectric; that way, the tunneling of electrons from the gate electrode into oxygen vacancies located within the HK dielectric becomes easier. The PBTI's new affection level is attributed to electron trapping into preexistent traps within HK dielectric layer, as well as trap generation due to electric and thermal stress while the device is under operation.

Just like NBTI, the PBTI model consists of three uncorrelated components: (1) interface traps' generation at the IL/HK interface, (2) electron trapping into preexistent defects within the HK dielectric layer, and (3) trap generation within the HK dielectric layer during stress. Just like equation (2.7), the equation (2.8) is given for PBTI. Figure 2.13 shows PBTI dynamics governed by interface traps' generation in HKMG technology. In such a case, interface trap generation at the interfacial and high-*k* dielectric layer interface, as well as activation of passivated oxygen vacancies within the HK dielectric are attributed to O_v-H breaking due to electron tunneling from the inversion channel, within the interfacial dielectric and towards the IL/HK interface. Released hydrogen at the IL/HK interface diffuses, reacts, and breaks passivated defects at the high-*k* dielectric and the metal gate electrode (HK/MG) interface. H_2 molecules diffusing from the HK/MG interface is the ruling component for PBTI. PBTI is treated as a newly-discovered degradation phenomenon because almost nothing is known about it.

$$\Delta V_T = \frac{q}{C_{ox}} \left(\Delta N_{IT-HK} + \Delta N_{ET-HK} + \Delta N_{OT-HK}\right) \tag{2.8}$$

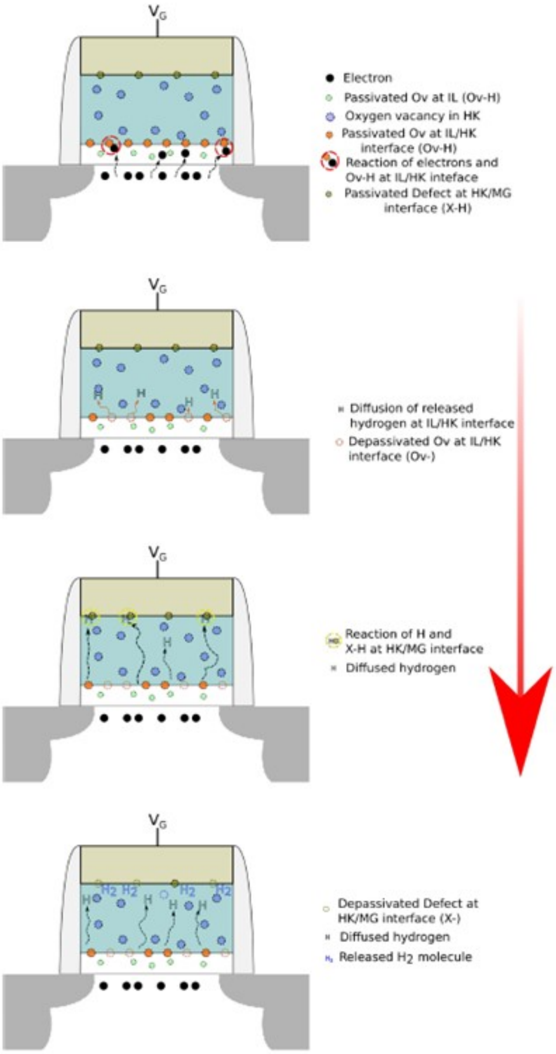

V_G
Electron
Passivated Ov at IL (Ov-H)
Oxygen vacancy in HK
Passivated Ov at IL/HK interface (Ov-H)
Reaction of electrons and Ov-H at IL/HK inteface
Passivated Defect at HK/MG interface (X-H)
V_G
H
Diffusion of released hydrogen at IL/HK interface
Depassivated Ov at IL/HK interface (Ov-)
V_G
H
Reaction of H and X-H at HK/MG interface
Diffused hydrogen
V_G
H2
H
Depassivated Defect at HK/MG interface (X-)
Diffused hydrogen
Released H2 molecule

Figure 2.13: Degradation time evolution for PBTI interface traps' generation.

Aging Effects on SRAM Cells' Performance

Two of the four SRAM cell transistors are under BTI stress at any given time regardless of the stored value [17]. As figure 2.14 shows, due to the cross-coupled structure of SRAM cell, when Q is "0," M2 and M3 are under BTI stress and when the value of Q is “1,” and M1 and M4 are under BTI stress and their relevant absolute V_T values increase. Unlike the combinational parts, the main effect of BTI on SRAM cells is SNM degradation rather than delay [18], [24].

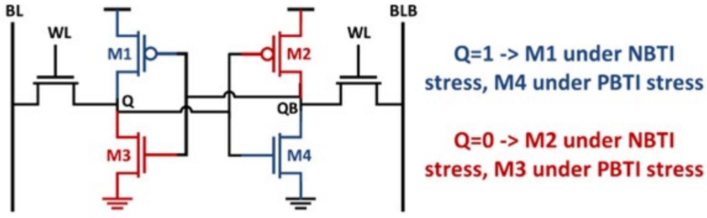

BL
WL
M1
M2
WL
BLB
Q
QB
M3
M4
Q=1 -> M1 under NBTI stress, M4 under PBTI stress
Q=0 -> M2 under NBTI stress, M3 under PBTI stress

Figure 2.14: SRAM cell structure.

The SNM is the stability indicator of SRAM cell and is defined as the minimum dc noise voltage required to flip the content of the SRAM cell [27]. The graphical description of SNM by using voltage transfer characteristic (VTC) curve of SRAM cell is the side length of the larger square that fits between two curves. The VTC curve of SRAM cell is called butterfly curve.

Based on SRAM operation modes, there are two definitions of SNM; HOLD, and READ SNMs [17]. The HOLD SNM is measured when word line is set to “0” and the SRAM cell is holding the data, and the READ SNM is measured when the word line is set to “1” and data is read from the cell. The HOLD SNM is always larger than READ SNM, also READ SNM is more susceptible to the V_T changes. This is because when the cell is holding its state, two inverters are strongly coupled to each other, and hence the HOLD SNM has less sensitivity to V_T shifts.

V_T change due to BTI shifts VTC curve, which results in the SNM degradation [28]. For instance, figure 2.15 shows the SNM degradation in the case of 50-, 100-, and 150-mV increase in V_T of p-type transistors for two cases: 1) when two p-type transistors have equal V_T shift and 2) when p-type transistors have unequal V_T shift, which means mentioned V_T shift is applied to only one p-type transistor. It can be seen that the SNM is reduced considerably by the increase of V_T shifts and gets worse when V_T shifts of transistors are asymmetrical.

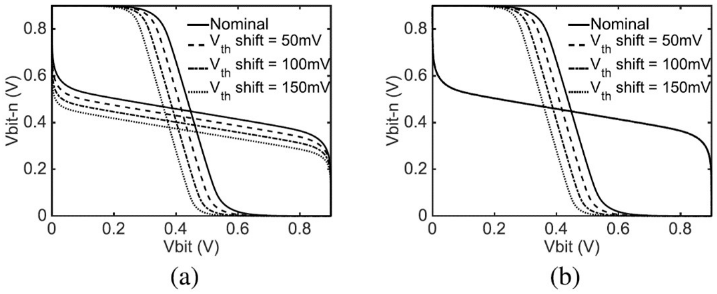


Figure 2.15: SNM reduction of SRAM cell over 50-, 100-, 150-mV V_T shift when p-type transistors are (a) symmetrically aged and (b) asymmetrically aged.

In the following of V_T increase, due to BTI, bit-flip can occur during the read operation. This is done when in read operation, storage node with “0” voltage rises to the trip point of its load inverter. Therefore, the content of SRAM cell will flip during the read operation and leads to destructive read.

2.2.2 PVT Variations

The continuous shrinking of transistor size leads to great advances in circuit performance besides reducing energy consumption and transistor cost. However, this aggressive scaling also makes the CMOS circuits more susceptible to variability. There exist 3 main sources of variability, namely, Process, Voltage and Temperature (PVT) variations. Process variations are due to the mismatch of the manufacturing process. Voltage variations are mostly due to the parasitic impedance while temperature variations are caused by the power dissipated by the circuit. PVT variations change the transistor switching speed and leakage current.

Process Variation

This variation accounts for deviations in the semiconductor fabrication process. Usually process variation is treated as a percentage variation in the performance calculation. Variations in the process parameters can be impurity concentration densities, oxide thicknesses and diffusion depths. These are caused by non-uniform conditions during depositions and/or during diffusions of the impurities. This introduces variations in the sheet resistance and transistor parameters such as threshold voltage. Variations are in the dimensions of the devices, mainly resulting from the limited resolution of the photolithographic process. This causes (W/L) variations in MOS transistors.

Process variations are due to variations in the manufacture conditions such as temperature, pressure and dopant concentrations. The ICs are produced in lots of 50 to 200 wafers with approximately 100 dice per wafer. The electrical properties in different lots can be very different. There are also slighter differences in each lot, even in a single manufactured chip. There are variations in the process parameter throughout a whole chip. As a consequence, the transistors have different transistor lengths throughout the chip. This makes the propagation delay to be different everywhere in a chip, because a smaller transistor is faster and therefore the propagation delay is smaller.

Supply Voltage Variation

The design's supply voltage can vary from the established ideal value during day-to-day operation. Often a complex calculation (using a shift in threshold voltages) is employed, but a simple linear scaling factor is also used for logic-level performance calculations.

The saturation current of a cell depends on the power supply. The delay of a cell is dependent on the saturation current. In this way, the power supply inflects the propagation delay of a cell. Throughout a chip, the power supply is not constant and hence the propagation delay varies in a chip. The voltage drop is due to nonzero resistance in the supply wires. A higher voltage makes a cell faster and hence the propagation delay is reduced. The decrease is exponential for a wide voltage range. The self-inductance of a supply line contributes also to a voltage drop. For example, when a transistor is switching to high, it takes a current to charge up the output load. This time varying current (for a short period of time) causes an opposite self-induced electromotive force. The amplitude of the voltage drop is given by the equation (2.9), where L is the self inductance and I is the current through the line.

$$V = L \times \frac{dI}{dt} \tag{2.9}$$

Temperature Variation

Temperature variation is unavoidable in the everyday operation of a design. Effects on performance caused by temperature fluctuations are most often handled as linear scaling effects, but some submicron silicon processes require nonlinear calculations.

When a chip is operating, the temperature can vary throughout the chip. This is due to the power dissipation in the MOS-transistors. The power consumption is mainly due to switching, short-circuit and leakage power consumption. The average switching power dissipation (approximately given by $P_{average} = C_{load} \times V_{powersupply} \times 2 \times f_{clock}$) is due to the required energy to charge up the parasitic and load capacitances. The short-circuit power dissipation is due to the finite rise and fall times. The NMOS and PMOS transistors may conduct for a short time during switching, forming a direct current from the power supply to the ground. The leakage power consumption is due to the nonzero reverse leakage and sub-threshold currents. The biggest contribution to the power consumption is the switching. The dissipated power will increase the surrounding temperature. The electron and hole mobility depend on the temperature. The mobility (in Si) decreases with increased temperature for temperatures above –50 ^{o}C. The temperature, when the mobility starts to decrease, depends on the doping concentration. A starting temperature at –50 ^{o}C is true for doping concentrations below 1019 atoms/cm^3. For higher doping concentrations, the starting temperature is higher. When the electrons and holes move slower, then the propagation delay increases. Hence, the propagation delay increases with increased temperature. There is also a temperature effect, which has not been considered. The threshold voltage of a transistor depends on the temperature. A higher temperature will decrease the threshold voltage. A lower threshold voltage means a higher current and therefore a better delay performance. This effect depends extremely on power supply, threshold voltage, load and input slope of a cell. There is a competition between the two effects and generally the mobility effect wins. Figure 2.16 shows the PVT operating conditions.

The best and worst design corners are defined as follows:

- Best case: fast process, highest voltage and lowest temperature.

- Worst case: slow process, lowest voltage and highest temperature.

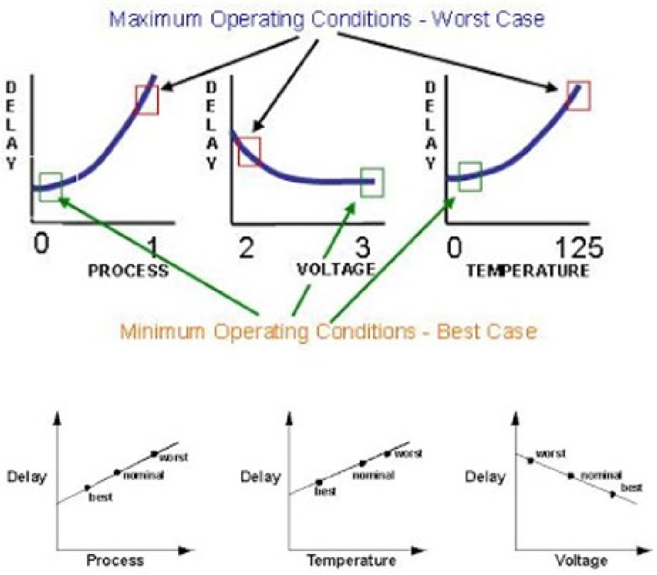


Figure 2.16: Delay performance for PVT variations.

2.3 Subthreshold Analysis

The Internet of Things (IoT) enables easy access and interaction with a wide variety of devices, some of them self-powered, consisting of microcontrollers, sensors and sensor networks. Therefore, it is important to use power management strategies and reduce power consumption in IoT chips. One of these techniques is the Dynamic Voltage and Frequency Scaling which allows energy consumption to be reduced at subthreshold power supply voltages. However, reducing the power supply voltage, implies the reduction of performance and, consequently, delay increases, which in turn makes the circuit more vulnerable to operational-induced delay-faults and transient-faults. For this reason, it is important to identify a compromise where power is drastically reduced, but most errors are still avoided or prevented.

2.3.1 Subthreshold Design Techniques

One of the most important areas regarding the optimization of energy for digital circuits is subthreshold design techniques. This area becomes important for circuits whose application does not require permanent and intensive performance, or for applications where processing speed is not a critical factor, such as digital circuits [9][10], analog circuits [5][6], mixed-signal applications, or even at memory applications [23][3].

Some works have already been carried out on this topic. In the works [7][8] the modeling and characterization of new devices designed specifically for operation at subthreshold levels is mentioned. Other works, such as [13]-[16], are focused on trying to establish ground rules and methods on how to design logic devices, that can fully work on optimum energy points at subthreshold modes. In [13][14] the concept of energy minimization is defined, and analytical methods are presented to allow calculating the optimum V_{DD} and V_T, for a specific operating frequency and minimizing power. Other work as in [15] refers to the importance of the size of transistors to optimize energy reduction in subthreshold circuits. The work [22] consists of design techniques that aim to minimize operational errors by introducing new fault tolerant methods, as well as new and more robust cell design techniques to significantly improve liability of digital circuits.

To determine the optimal operating conditions, there are works [9][10] that feature the complete design of a new standard cell library fitted to work at subthreshold voltage levels. However, to define an optimal V_{DD} value for an ultra-low power operation it is necessary to take into account the compromise between several parameters and different gates. Moreover, reusing an existing standard cell library to work at subthreshold voltages can lead to good results.

2.3.2 Energy Savings Techniques

One method of obtaining energy savings is to use the technique Dynamic Voltage and Frequency Scaling (DVFS), which allows to dynamically change the supply voltage and frequency during the operation of a given task of a circuit.

Because a processor's workload is not constant, it means that there are times when the processor needs more power and others when it needs less power (idle moments). So instead of the processor working at a constant supply voltage, we can vary it over time so that in idle moments the processor consumes less power supply, saving energy. Figure 2.17 shows that to perform two tasks, instead of using 100% of the supply voltage, we can use a lower supply voltage over a longer time interval, thus reducing idle moments and saving more energy.

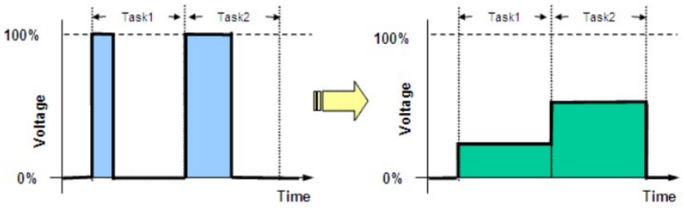


Figure 2.17: DVFS main principle.

Another way to save energy with this technique is to change the operating frequency. The power dissipated for common digital CMOS circuits is given by the following equation:

$$P_{Total} = C \times V_{DD}^2 \times f + (I_{sub} + I_{diode} + I_{gate}) \times V_{DD} \tag{2.10}$$

Through this equation it is possible to verify that the dissipated power is the sum of two terms. The first term corresponds to the dynamic power dissipation that depends on frequency and supply voltage. Thus, to decrease power dissipation, it is only necessary to decrease the operating frequency and supply voltage at idle moments. The second term of the equation is static power dissipation which is related to various leakage currents occurring on the circuit and may generally be considered much lower than the first term.

The technique defined as DVFS intends to maintain performance and guarantee system's reliability at all times, while obtaining energy savings. To achieve it, the technique works by dynamically reducing frequency along with supply voltage, thus compensating critical path increase. According to this method, the system's main features such as supply voltage, frequency and critical path must be carefully monitored and permanently adjusted, so that energy consumed can be minimized, while maintaining system's performance as required.

2.3.3 The Compromise Between Power and Performance

In the subthreshold analysis, there is a compromise between power and performance, that is, if the power decreases considerably, the performance also decreases, but if an increase in performance is required, it is not possible to have minimal power consumption. This implies that the optimal value for V_{DD}, will not be the minimum value, but rather the best compromise between power and performance, reducing considerably power but not jeopardizing performance, neither the correct operation with an increased vulnerability to errors.

To measure the efficiency of performing an operation in a given technology, a figure of merit like power-delay product (PDP) is used. As power times delay has the dimension of energy, this figure of merit is also known as the switching energy, because it is the product of the average power consumption over a switching event times the input-output propagation delay of the event, or duration of switching event ($Power \times Delay^2$). Minimizing the PDP of a circuit results in a particular design point in the energy-delay space where 1% of energy can be traded off for 1% of delay.

There are also other metrics similar to PDP in which the delay assumes greater weight as is the case of $Power \times Delay^n$ and even metrics that use sensitivity [29].

Chapter 3

State of the Art on Performance Sensors

As mentioned in the previous chapter, PVTA variations can result in circuit degradation and consequently failures in RAM memories during their various states of operation. With this in mind, this chapter presents some work done on aging and performance sensors for cells of SRAM and DRAM memories. Unfortunately, there are not many studies on the subject and will be presented only aging sensors caused by the NBTI effect (OCAS), performance sensors for SRAM and an improved version compatible with DRAM memories (Scout Memory Sensor), and finally a sensor for logic circuits (LPS and GPS).

3.1 On-Chip Aging Sensor (OCAS)

On-chip aging sensor (OCAS) is a sensor that permits to detect SRAM aging caused by NBTI effect during system lifetime. The sensor is able to detect any specific aging state of a cell in the SRAM array. The strategy is based on the connection of an OCAS per SRAM column, which periodically performs off-line testing by monitoring write operations into the SRAM cells to detect aging. This approach is application-transparent since it is does not change the SRAM contents after testing. To prevent OCAS from aging by one side and from dissipating static power by the other side, OCAS circuitry is powered-off during idle periods.

In figure 3.1 is shown the general block diagram of the proposed approach indicating the connection between the OCAS and one SRAM column and figure 3.2 shows the OCAS's schematic.

As observed, transistor TT1 is connected between the real V_{DD} and virtual V_{DD} node (V'_{DD}), which is used to feed the positive bias to the cells of the SRAM column. During Normal Operating Mode, TT1 is on, while the OCAS is powered off by the p-type (TPG) and n-type (TNG) transistors shown in figure 3.2. The Power Gating Technique (PGT) is used to switch transistors TPG and TNG off during the Normal Operating Mode, thereby any aging of the OCAS circuitry is avoided. During the Testing Mode, the OCAS is powered on by TPG and TNG, which both are turned on while TT1 is switched off. At this moment, a write operation, or a sequence of write operations, is performed on a specific memory cell

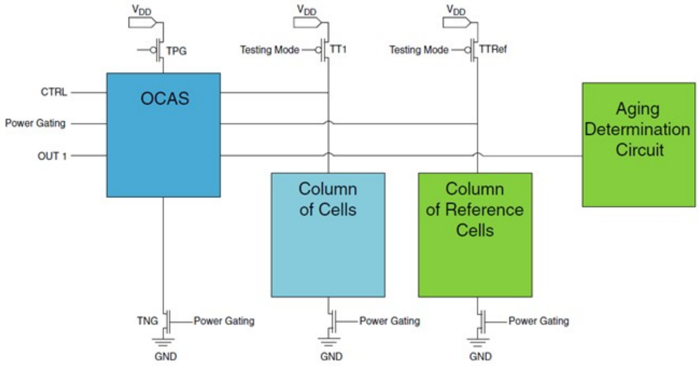


Figure 3.1: General block diagram of the hardware-based approach connected to one SRAM cell column.

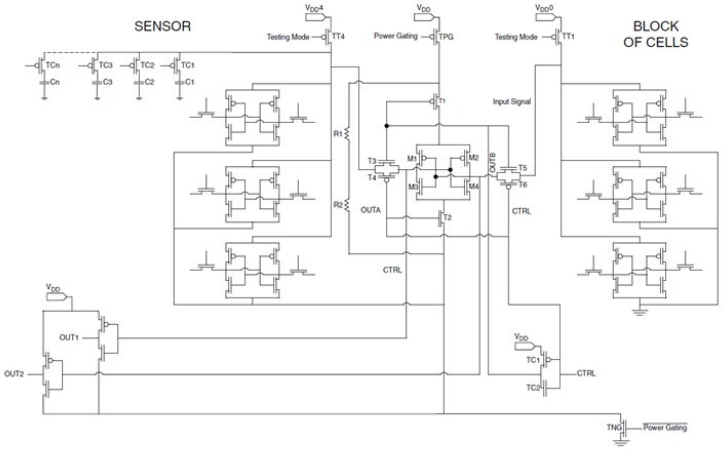


Figure 3.2: Schematic of the OCAS connected to a six-transistor cell column.

in order to measure its aging state. After performing a comparison between the V'_{DD} node's voltage at the end of a write operation and the Reference Voltage value previously adjusted inside the sensor, the OCAS takes the pass/fail decision. At the end of this process, the OCAS's output (OUT1) yields a logic “0” for a fault-free or new SRAM cell or a logic “1”, which represents a fault state or, in other words, that the cell is no more reliable due to its advance Aging state.

Observing figure 3.2 it is possible to see that the control signal CTRL is set to “0” during the pre-charge phase of the Testing Mode, whereas during the evaluation phase, this signal is set to “1”. Upon the pre-charge phase, transistors TC1, T3, T4, T5, and T6 are driven by the CRTL signal to the on state and the signals to be checked are driven into the voltage comparator formed by the transistors M1, M2, M3, and M4. In sequence, during the evaluation phase, CTRL is set to “1” turning TC1 as well as T3, T4, T5, and T6 off, while switching TC2, T1, and T2 on, which allows M1, M2, M3, and M4 to evaluate the input signal, the voltage at V'_{DD}, against the Reference Voltage value generated by resistors R1 and R2. If the sensing value coming from V'_{DD} is lower than the Reference Voltage, the cell is still categorized as non-aged; otherwise, the cell is considered aged and the OCAS output (OUT1) is set to “1”. This process is executed for each cell in the SRAM, of which one desires to measure the aging state.

It is important to mention that there is a small circuitry embedded in the OCAS, which is used to perform the sensor's self-test before it is being activated to monitor the SRAM cells. This small circuit, not shown in figure 3.2, consists of two resistors (R3 and R4 in series), which carry the same configuration as the resistors R1 and R2, but differently are connected to the drain of transistors M2 and M4. The voltage produced at this node is slightly smaller than the one produced at the V'_{DD} after a sequence of two write operations in an aged cell activated during the Testing Mode. As logical consequence, when activating the OCAS self-test, it is expected that OUT1 will indicate an error and therefore setting the logic level “1”.

Figure 3.3 summarizes the complete flow adopted in order to measure the aging state of SRAM cells.

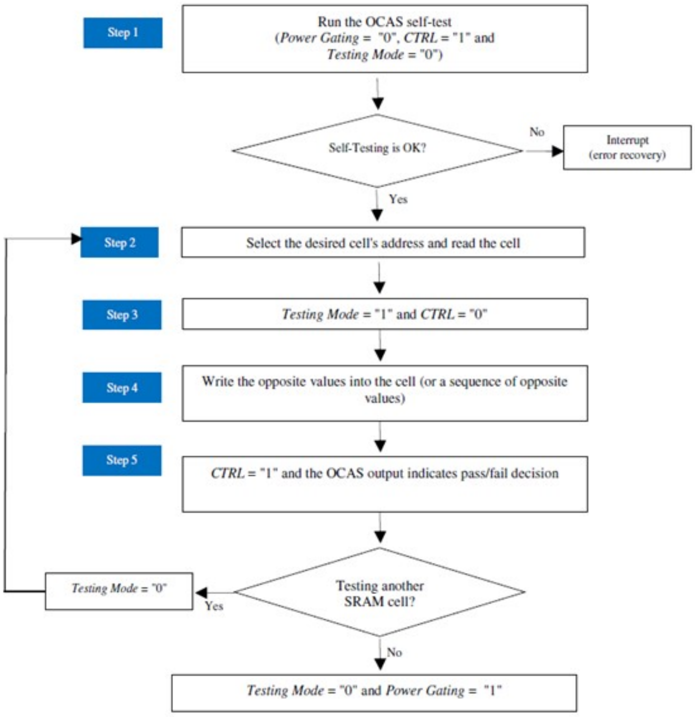

Step 1
Run the OCAS self-test (*Power Gating* = "0", *CTRL* = "1" and *Testing Mode* = "0")
Self-Testing is OK?
No
Interrupt (error recovery)
Yes
Step 2
Select the desired cell's address and read the cell
Step 3
Testing Mode = "1" and *CTRL* = "0"
Step 4
Write the opposite values into the cell (or a sequence of opposite values)
Step 5
CTRL = "1" and the OCAS output indicates pass/fail decision
Testing another SRAM cell?
Yes
Testing Mode = "0"
No
Testing Mode = "0" and *Power Gating* = "1"

Figure 3.3: Measurement flow adopted by the hardware-based approach.

3.2 Aging and Performance Sensor for SRAM

One of the most important blocks of RAM memories is the sense amplifier that allows you to detect small differences on the bit lines and the reestablishment of digital signals, by correctly reading stored values. This reading by the sense amplifier has a certain response time corresponding to the transition times of the bit line. When the circuits of memories and sense amplifier are new these transitions are fast, but when transistors ageing the transitions become slower due to the degradation of physical properties as illustrated in figure 3.4. Therefore, by monitoring the response time of a cell and measuring the switching times of the bit line signals it is possible to measure memory cells performance and, consequently allows aging monitoring. Taking this into account, an aging and performance sensor that detects errors when slow transitions occur due to aging effects during write and read operations for SRAM memories has been proposed in previous work.

This sensor is mainly constituted by two blocks as can be seen in figure 3.5. The first is the transition

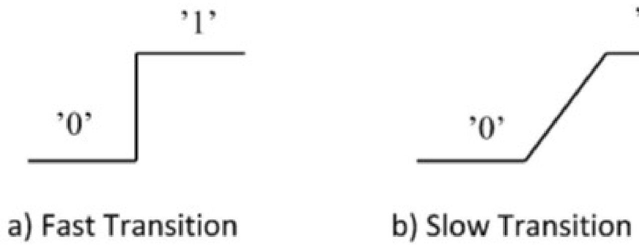


Figure 3.4: Transitions: a) Fast transition b) Slow transition.

detector, described in more detail in the following section, which generates pulses in the presence of a signal transition, on the memory cell bit lines and the second block is the pulse detector that indicates if the generated pulse (which has a duration proportional to the transition time) exceeds a defined value in the pulse duration, indicating a slow transition and, consequently a critical performance of the memory cell that could lead to a fault. In this case, an error output is generated.

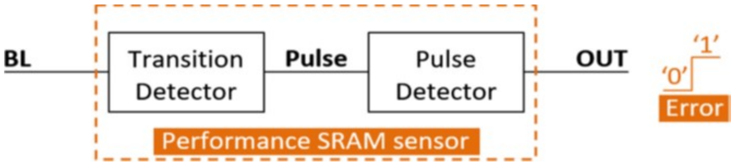


Figure 3.5: Aging and performance sensor block diagram.

3.2.1 Transition Detector

The architecture of the transition detector block is represented in figure 3.6.

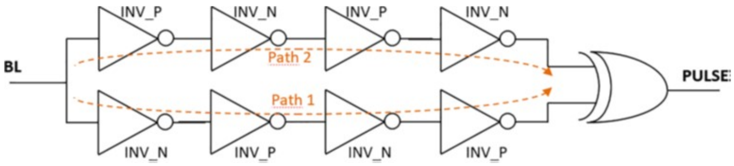


Figure 3.6: Transition detector's implementation.

Looking at figure 3.6, we can see that the transition detector is constituted by two paths, each with 4 inverters. In these 4 inverters there are 2 inverters with a more conductive NMOS MOSFET and other 2 inverters with a more conductive PMOS MOSFET. This leads to transitions from the bit line from "0" to "1", where path 2 is a faster transition and path 1 is a slower transition (figure 3.7).

While for bit line transitions from "1" to "0", path 2 corresponds to a slower transition and path 1 corresponds to a faster transition (figure 3.8).

To implement the inverters of these two paths, the following transistor sizes (table 3.1) were used for 65nm CMOS technology.

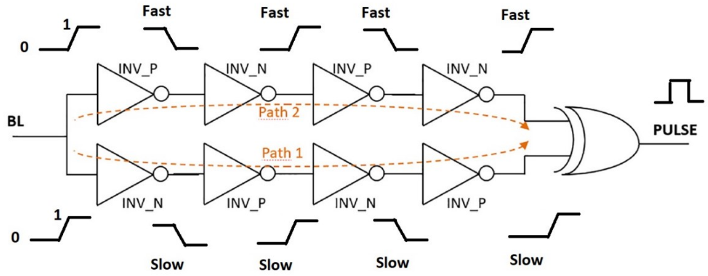


Figure 3.7: Transition detector for bit line transitions of ”0” to ”1”.

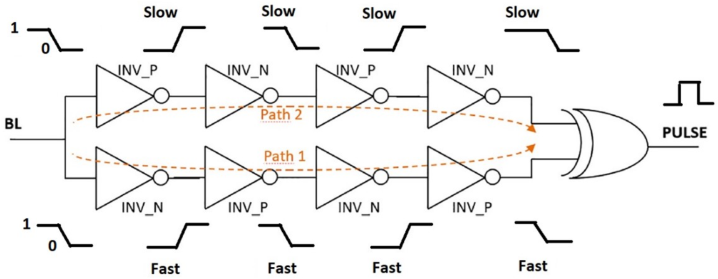


Figure 3.8: Transition detector for bit line transitions of ”1” to ”0”.

These two paths will link to a XOR gate that is not a classic CMOS XOR gate, but rather a pass-transistor logic XOR gate, which includes an inverter as its output, and ensures good performance without logic levels degradation (figure 3.9). This XOR aims to generate a pulse with a duration proportional to the transition time in the bit line.

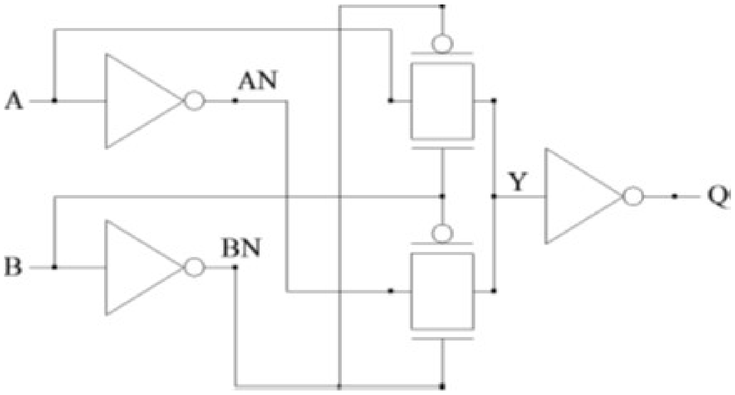


Figure 3.9: Pass-transistor XOR gate implementation.

Path	Inverter	NMOS	PMOS	L	$V_{T,n}$	$V_{T,p}$
1	Xinv1	5xWNmin	WPmin	65 nm	0,423 V	-0,365 V
	Xinv2	WNmin	5xWPmin			
	Xinv3	5xWNmin	WPmin			
	Xinv4	WNmin	5xWPmin			
2	Xinv1	WNmin	5xWPmin			
	Xinv2	5xWNmin	WPmin			
	Xinv3	WNmin	5xWPmin			
	Xinv4	5xWNmin	WPmin			

Table 3.1: Transition detector transistors’ sizes.

3.2.2 Pulse Detector

The pulse detector aims to indicate whether the pulse generated by the transition detector exceeds a certain value set by the clock, if this happens an error output is generated. There are two implementations for the pulse detector that are similar but have different modes of operations.

Stability-Checker Implementation

The first implementation for the pulse detector is the stability checker based on the Scout Flip-Flop, which detects all transitions in the data input that reaches the stability checker during the active pulse of the clock. This solution is robust and improves the sensitivity of the pulse detector in the presence of PVTA variations, because the delays generated by the circuit are also sensitive to PVTA variations and because the performance of the circuit is directly related to the frequency of the clock, i.e., if the clock frequency is reduced (increased), the performance is relaxed (excited) and the error probability is alleviated (aggravated).

Figure 3.10 shows that this implementation consists of a delay element, an inverter, and a stability checker. The delay element is basically a buffer, to provide a time delay to the input signal, and its architecture is presented in [23][3][4]. More than one delay element can be used, depending on the delay time required and the frequency of the clock. The stability checker is used here to detect transitions in the delayed pulses obtained from the delay element, but because the clock is connected to an inverter, it means that the stability checker detects transitions during the low state of the clock.

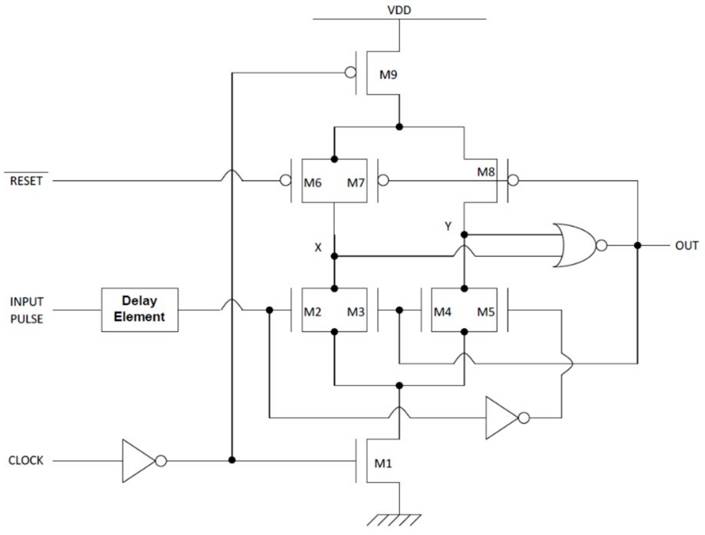


Figure 3.10: Stability-checker implementation.

The operation of this implementation requires a clock that is synchronized with all control signals and all memory instructions. Therefore, bit line transitions and pulses generated by the transition detector occur in the active state of the clock. These pulses undergo a propagation delay due to PVTA variations that are further accentuated by the presence of delay element. When this propagation delay to reach the stability checker during the low state of the clock, an error signal will be generated. This means that by design we have two parameters where we can control the delays in the sensor and the error/non-error decision: one is the sensibility of the transition detector and the width of the pulses generated; the other one is the time delay introduced in the signal in the delay element. In figure 3.11 is shown the stability checker operation for several signals. Through the figure it is possible to see that the grey region corresponds to slower transitions of the bit line signal indicating the performance reduction. In these cases, the pulses generated by the transition detector are wider, the delay added in the delay element is larger, which results in an error signal at the output of the sensor.

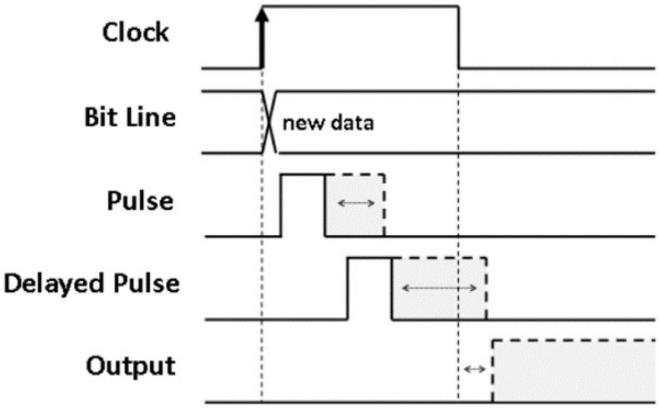


Figure 3.11: Pulse detector with stability-checker operation.

NOR-Based Pulse Detector Implementation

The second implementation is an improved version of stability checker, but this time it resorted to using a NOR to detect when simultaneously two signals are at low state. Figure 3.12 shows the architecture of this new implementation, which consists of 4 transistors that form a CMOS NOR logic gate (M1, M2, M4 and M5), controlled by a clock signal (CLOCK) and delayed pulses (Delayed Pulse). The inverter and transistors M3 and M6 ensure, in case of a detection, that the output signal (OUT) remains active until a reset occur (this allows to exempt the use of a latch to keep the sensor active in case of an error). The reset signal (RESET) controls M7 transistor operation and reinitiates all the circuit for new detection.

The operation of the NOR-Based pulse detector is basically the same as the stability checker, i.e. a clock signal is again used as a fixed reference to detect abnormal delays in the pulses generated by the transition detector. Considering that signals in the memory are generated in the rising edge of the clock, the pulses in pulse detector's input will also occur during the high state of the clock, but this time the pulses are reversed so that an error occurs when both the clock and the delayed pulses are in the low state, as can be seen in figure 3.13.

The main advantages are that this new implementation is less complex, reducing from 17 to 11 transistors and generating a smaller sensor area relative to stability checker. Regarding reliability, the stability checker has an intrinsic delay which becomes prohibitive when V_{DD} is reduced, making this solution improper to work with DVFS. On the contrary, the NOR-Based pulse detector is much more stable and reliable when working at reduced power-supply voltages (or even sub-threshold voltages),

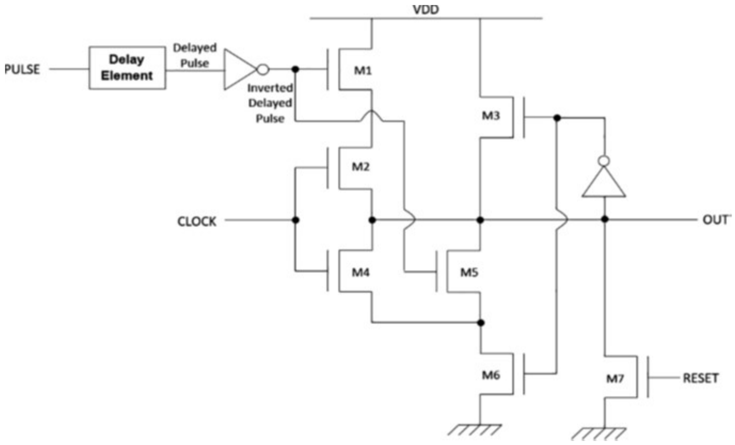


Figure 3.12: NOR-Based pulse detector implementation.

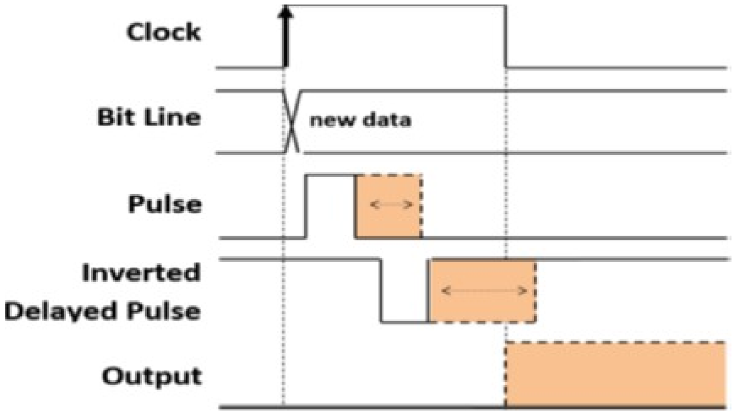


Figure 3.13: NOR-Based pulse detector operation.

because of the simpler behavior based on the OR gate. Regarding power, the previous work uses dynamic CMOS logic, which imposes constant switches of signals in every clock cycle. This behavior imposes higher dynamic power dissipation when compared with the classic CMOS OR gate behavior of the new pulse detector.

However, this new implementation still has some disadvantages because this sensor does not apply to DRAM memories, nor does it allow the user to change sensitivity or calibrate it making it more versatile. As explained earlier, this sensor needs to be synchronized with memory, which represents a major limitation for this type of sensor.

3.3 Scout Memory Sensor

In this section is presented a sensor (Scout Memory Sensor) that detects degradation of memory circuits caused by PVTA variations, allowing to avoid the occurrence of errors during read and write operations, signaling appropriately when the performance of memories is at risk. The novelty of this sensor is that it allows its use in SRAM and DRAM memories and also allows the user to calibrate and change the sensitivity of the sensor, making this solution more versatile and solid. Another advantage of this sensor is that it becomes more sensitive if the signal degradation is greater, i.e., its sensitivity is improved when operating conditions are worse. This sensor can be used as a global sensor (monitoring of all memory) or as a local sensor (monitoring of a specific location). Another advantage that this sensor presents is that it can work online, during the normal circuit operation, without the need to go offline.

The Scout Memory Sensor consists of 4 blocks (figure 3.14) and also features a controller that functions as a Finite State Machine (FSM) with 3 states, *Reset*, *Sample* and *Compare* that allow the control and operation of the entire sensor. The 4 sensor blocks are: the transition detector, the pulse detector, the reference value for comparison and the comparator. The transition detector is connected to the bit line and generates a pulse for each transition that occurs on the bit line. This pulse has a duration proportional to the transition time of the bit line. The pulse detector aims to generate a DC voltage proportional to the pulse duration generated by the transition detector. This pulse detector has a system that allows the user to control sensitivity via 3-bit control. The reference value block creates a reference voltage close to V_{DD} ($V_{DD} - V_T$), for comparison purposes. The last block (comparator) compares the reference voltage with the DC voltage obtained in the pulse detector.

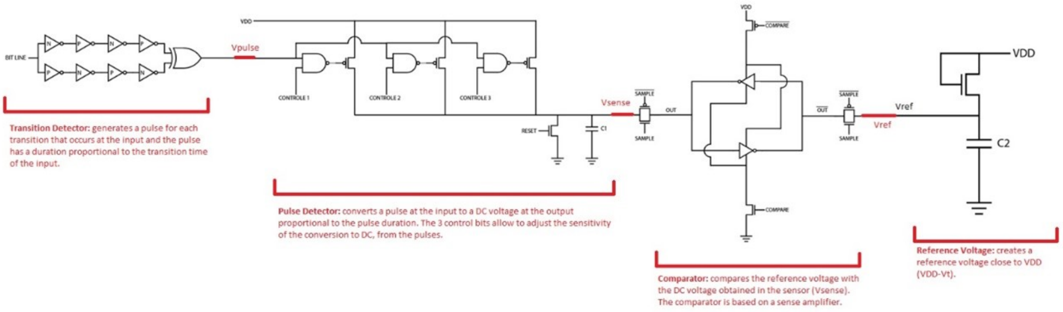

BIT LINE
N
P
N
P
P
N
P
N
VDD
Vpulse
CONTROLE 1
CONTROLE 2
CONTROLE 3
RESET
C1
Vsense
$\overline{\text{SAMPLE}}$
SAMPLE
OUT
VDD
$\overline{\text{COMPARE}}$
COMPARE
$\overline{\text{OUT}}$
$\overline{\text{SAMPLE}}$
SAMPLE
Vref
Vref
VDD
C2
Transition Detector: generates a pulse for each transition that occurs at the input and the pulse has a duration proportional to the transition time of the input.
Pulse Detector: converts a pulse at the input to a DC voltage at the output proportional to the pulse duration. The 3 control bits allow to adjust the sensitivity of the conversion to DC, from the pulses.
Comparator: compares the reference voltage with the DC voltage obtained in the sensor (Vsense). The comparator is based on a sense amplifier.
Reference Voltage: creates a reference voltage close to VDD (VDD-Vt).

Figure 3.14: Scout Memory Sensor architecture.

3.3.1 Transition Detector

The Scout Memory Sensor transition detector is basically the same as it was used in the previous sensor and works the same way, i.e. the transition detector receives a bit line transition and generates a pulse (Vpulse) that is proportional to the duration of the transition, and faster transitions generate pulses with smaller width and slower transitions generate pulses with greater width. Slower transitions reflect the degradation of memory circuits caused by PVTA variations.

In figure 3.15, it is possible to observe the structure of the transition detector, which is again composed of two paths of 4 inverters each and that will connect to an XOR gate. These 4 inverters have different conductivities which leads to a faster transition to the top path and a slower transition to the bottom path when the transition from the bit line is from ”Low” to ”High” and for transitions from ”High” to ”Low”, the top path has slower transitions and the low path transitions faster.

The process of operation of this block is exactly the same and can be seen in section 3.2.1.

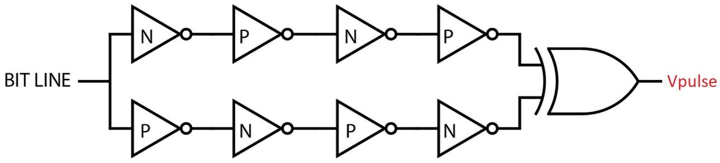


Figure 3.15: Transition detector for Scout Memory Sensor.

3.3.2 Pulse Detector

The Scout Memory Sensor pulse detector is a block that is different from the others used previously. The basic idea of this block is to receive a pulse (Vpulse) from the transition detector and convert it into a DC voltage to the output with a value proportional to the pulse duration, that is, for pulses with longer duration, the DC voltage will have a higher value and for pulses with shorter duration, the DC voltage will be lower. This is achieved through the time it takes to charge a capacitor (the shorter the pulse duration, the less time it takes the capacitor to charge, which implies a lower DC voltage on the output).

In figure 3.16 we can see that the pulse detector consists of 3 NAND gates, 3 PMOS transistors, an NMOS transistor controlled by *Reset* state that ensures that the sensor's start-up conditions are always the same as when a new test is carried out and a capacitor (C1) that generates a DC value (Vsense) proportional to the time it takes to charge.

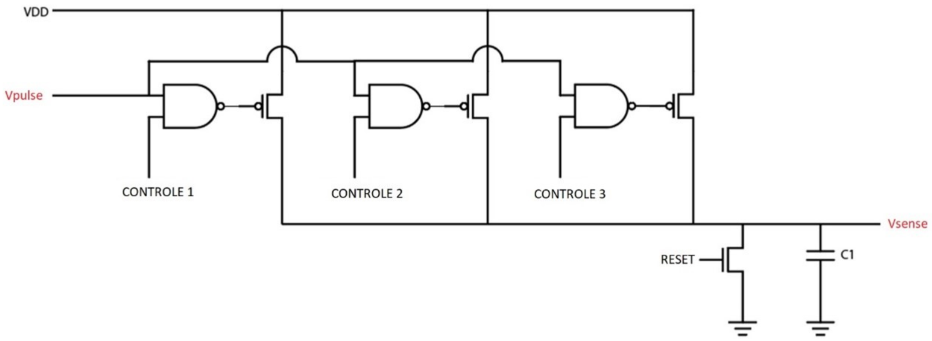


Figure 3.16: Pulse detector for Scout Memory Sensor.

This sensor presents a novelty compared to previous work, because through 3-bit control allows the user to change the sensitivity of the sensor during online operation. The 3 control bits have 7 sensitivity levels that allow you to change the current that will charge the C1 capacitor, increasing or decreasing the speed with which the capacitor loads (for example, if the 3 PMOS transistors are conducting, the capacitor charges faster and the Vsense value will be higher). It may be possible to increase sensitivity levels, but this will have a higher hardware cost and consequently larger area (plus NAND gates and PMOS transistors). Table 3.2 shows the size of PMOS transistors that allow the user to change the sensor sensitivity. Another way to change the sensitivity of the sensor is to change the capacity of the C1 capacitor.

Path	Size PMOS	L	$V_{T,n}$	$V_{T,p}$
CONTROLE 1	WNmin	65 nm	0,423 V	-0,365 V
CONTROLE 2	2xWNmin			
CONTROLE 3	4xWNmin			

Table 3.2: Pulse Detector - Size of the transistors.

3.3.3 Reference Value for Comparison

The third block is the Reference Value with the aim of creating a reference voltage (Vref), which will be used by the comparator block to compare with Vsense.

From figure 3.17 we can observe that this block consists of an NMOS transistor connected to a C2 capacitor.

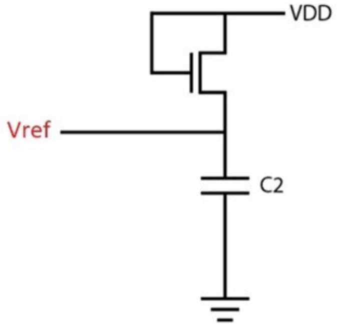


Figure 3.17: Reference value for Scout Memory Sensor.

Because both the drain and gate of the NMOS transistor are connected to V_{DD}, this implies that the transistor is in the saturation region and conducts a current with the following expression:

$$I_D = \frac{1}{2}\mu_n C_{ox}\left(\frac{W}{L}\right)(V_{GS} - V_{T,n})^2 \tag{3.1}$$

In turn this current will charge a C2 capacitor until to a total load of $V_{DD} - V_{T,n}$ that corresponds to the reference value (Vref), to be compared by the comparator.

As the NMOS transistor of this block ages more than memory, which implies a lower Vref voltage due to increased $V_{T,n}$, this means that the higher the aging degradation of the circuits, the greater the sensitivity of this sensor, which constitutes a great advantage of this type of sensor.

3.3.4 Comparator

The last block of the Scout Memory Sensor is the comparator that serves to compare the reference voltage (Vref) with the DC voltage obtained in the pulse detector (Vsense).

Through the figure 3.18, it is possible to observe that the comparator is constituted by two transmission gates that are controlled by the *Sample* signal. When this signal is activated, the Vsense and Vref voltages are sampled and will be used by the comparator for comparison purposes. The comparator is also constituted by two cross-coupled inverters and an NMOS and PMOS transistor that are controlled by the *Compare* signal. When this signal is activated the two cross-coupled inverters are turned on and the comparison begins.

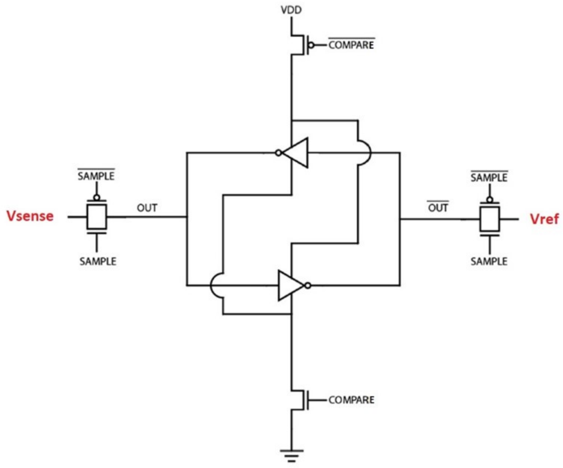

VDD
$\overline{\text{COMPARE}}$
$\overline{\text{SAMPLE}}$
Vsense
OUT
SAMPLE
$\overline{\text{SAMPLE}}$
$\overline{\text{OUT}}$
Vref
SAMPLE
COMPARE

Figure 3.18: Comparator for Scout Memory Sensor.

This comparator functions as a sense amplifier (explained in chapter 2), that is, it receives the sampled vsense and Vref signals and amplifies the differences between these two signals. When the Vref signal is greater than Vsense, the output of the sensor (OUT) is V_{SS}, indicating that the transition of the bit line is fast enough that no errors occur during write and read operations on the memory. But when the Vref signal is less than Vsense, the output OUT is V_{DD}, indicating that the transition of the bit line is too slow and that we are in the event of error due to PVTA effects that memory suffers.

3.3.5 Controller

This sensor also features a controller that is basically the Finite State Machine (FSM) that allows you to generate the signals *Reset*, *Sample* and *Compare* and control all sensor operation. To generate these 3 signals, figure 3.19 shows a state diagram.

As can be seen from the figure, this diagram consists of 3 states. The first is the *Reset* that serves to reset the entire sensor operation to an initial state, so that no errors occur in the measurements. The second state is the *Sample* that serves to place the FSM waiting for a transition to happen in the bit line. When this transition occurs, a pulse is generated and the FSM moves to the next state (*Compare*), otherwise it remains in the same state until a pulse is detected. In the *Compare* state, the Vsense and Vref signals are compared. The FSM will remain in this state, until read and writing of the memory occurs. When a new reading and writing of memory occurs the FSM returns to the initial reset state and there is a pre-charge signal that is activated to trigger the initialization of the bit lines.

The implementation of the controller circuit can be done using the Karnaugh map and is shown in figure 3.20, where two D flip-flops were used.

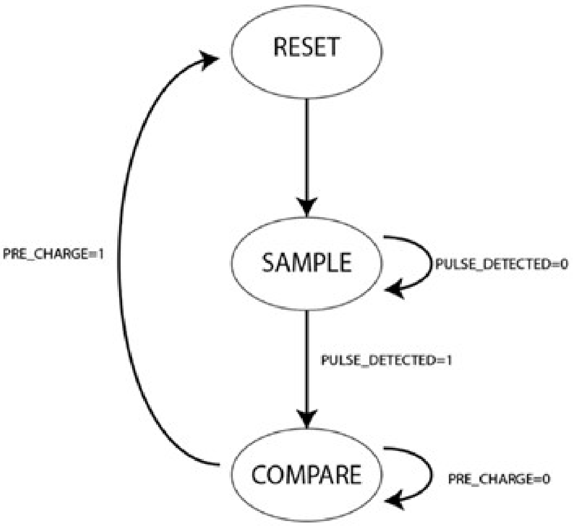


Figure 3.19: State Machine of the sense performance.

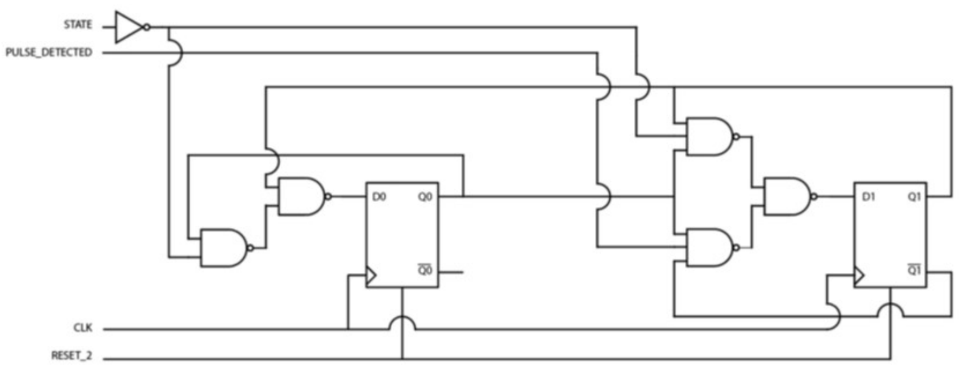


Figure 3.20: Controller for Scout Memory Sensor.

Figure 3.21 allows to see the signals coming from the flip-flops of the FSM, which indicate the state of the FSM, and the logic gates that generate the sensor's control signals from the state of the FSM.

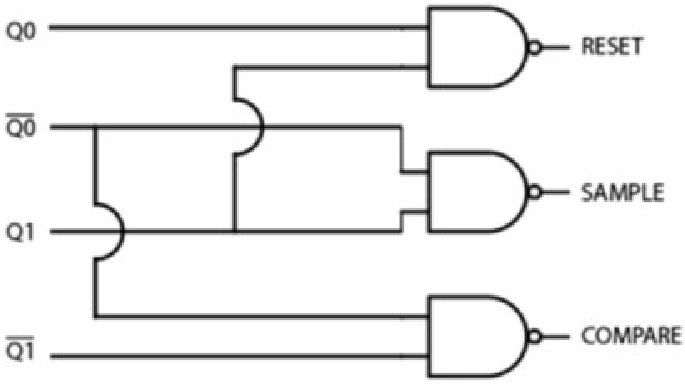


Figure 3.21: Control signals of the sensor.

Note also that to generate the *State* and *Pulse_Detected* signals, a D latch with *Sample* input and a clock *Pulse_In* are used (figure 3.22). When the *Pulse_In* is "High" the *Sample* appears on the latch output and when *Reset* is activated the latch output is ”Low”.

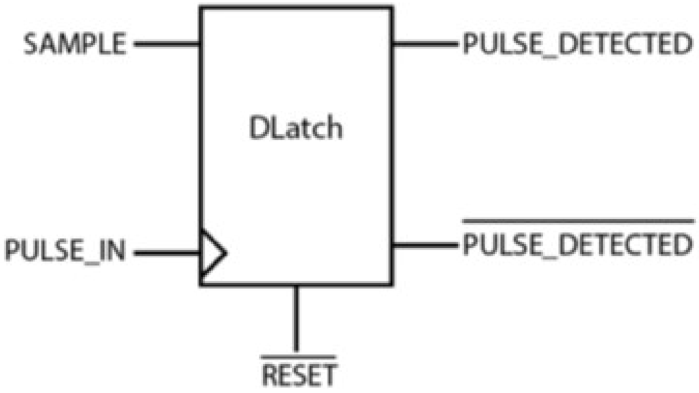

SAMPLE
PULSE_DETECTED
DLatch
PULSE_IN
$\overline{\text{PULSE\_DETECTED}}$
$\overline{\text{RESET}}$

Figure 3.22: Pulse detection latch.

3.4 Scout Performance Sensor for Synchronous Logic Circuits

There are aging sensors for synchronous logic circuits, which can work at subthreshold voltages to considerably reduce power consumption. This type of performance sensor utilizes an adaptive voltage scaling (AVS) strategy to enhance reliability and fault-tolerance and allows circuits to be dynamically optimized during their lifetime while preventing error occurrence.

In this section is made the approach of two types of sensors for synchronous circuits: the local performance sensor (LPS) to monitor performance degradations locally, in the actual circuit implementing the mission functionality, in key locations in the circuit where errors are more prone to occur. However, their implementation in a circuit is more complex and performance monitoring can only be done on-line if, and when, the critical paths they monitor are activated, which depends on circuit operation. The other sensor is the global performance sensor (GPS) to monitor key critical paths, critical paths' replicas, or key parameters, to detect performance degradation. Their usage in a circuit is very easy and straight forward, because performance monitoring, normally, is independent from circuit operation, which is why they are easily adopted by industry. However, they do not monitor circuit at the actual locations where error occur, and, because of that, their estimated Process, Voltage, Temperature and Aging (PVTA) variations may differ from the ones that in the real circuit can produce an error.

Finally, a complete performance sensor for subthreshold operation is presented, consisting of LPS and GPS that allows an ultra-low-power strategy for reliable IoE nanoscale digital circuits.

3.4.1 Local Performance Sensor

The local performance sensor (LPS) for subthreshold operation is a sensor that allows to monitor data transitions at key flip-flops (FF), in order to identify the occurrence of unsafe transitions. To determine locally these unsafe transitions, delays are used inside the FF to create virtual windows and detect these unsafe transitions (i.e., data transitions which are in the eminence of causing a delay-fault). The detection of unsafe transitions and the capture of late transients, allow to monitor and control circuit operation when working with a high variability in circuit performance (namely, using reduced V_{DD} voltages, or using AVS techniques).

The LPS architecture is shown in figure 3.23.

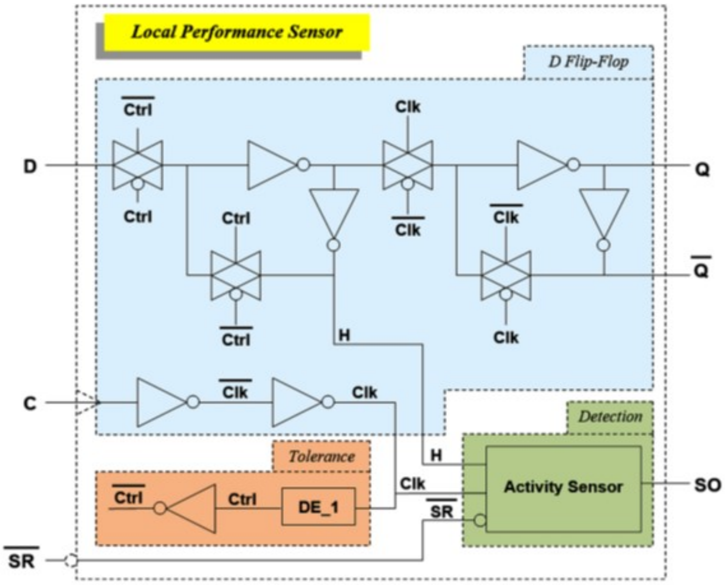


Figure 3.23: Local performance sensor architecture.

This sensor consists of three blocks. The first block is an D-type flip-flop, which include a common master-slave D FF with a data input D, a clock input C, and the data outputs Q and $\overline{Q}$. The second block is the delay-fault tolerance, which include two additional internal signals, $Ctrl$ and $\overline{Ctrl}$, to generate a delayed clock signal in the master latch and to provide an additional time to capture late transients in the FF. The third block is the detecting unsafe data transitions, which include an activity sensor block, to signalize transitions in the eminence of an error in the internal data signal H, an additional sensor output signal, SO, and an additional sensor reset signal, $\overline{SR}$.

To better understand the operation of the LPS, figure 3.24 shows the timings and delay margins in key signals in the LPS, in respect to the clock period.

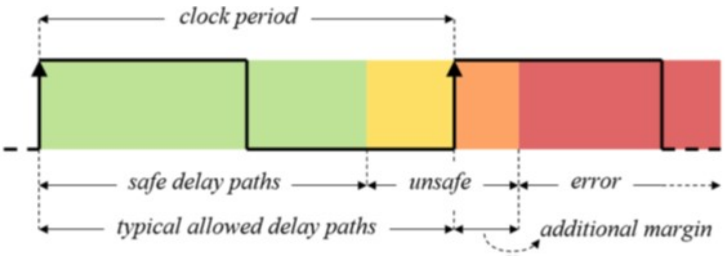


Figure 3.24: Path-delay margins in the LPS within the clock period.

Through the figure it is possible to see that in a typical flip-flop, the allowed delay in a data path is, utterly, the clock period. There is a margin time, less than a clock period, where signals can reach the FF input, without the occurrence of error (safe margin). There is also a time margin, where FF captures data correctly but in the eminence of an error (unsafe margin). This unsafe margin can be longer than a clock period and is signaled by the sensor in order to take steps to avoid errors. If late transients occur beyond the unsafe margin, an error data signal is captured in the FF.

As mentioned earlier, the activity sensor block aims at the detection of unsafe data transitions at the FF input and was designed to work at reduced V_{DD} voltages levels, down to subthreshold levels. Figure 3.25 shows the architecture of the activity sensor block and it is possible to observe that on the output of the XOR gate is generated the pulse (det signal) for every transition in H signal, with its pulse duration being proportional to the propagation delay of the DE_2 block. The sensor output (SO signal) is signalized (output high) when the generated pulse (det signal) and the clock (*Clk*) are simultaneously active (high). This sensor has the advantage of being more sensitive when operating conditions worsen, i.e. when PVTA variations increase.

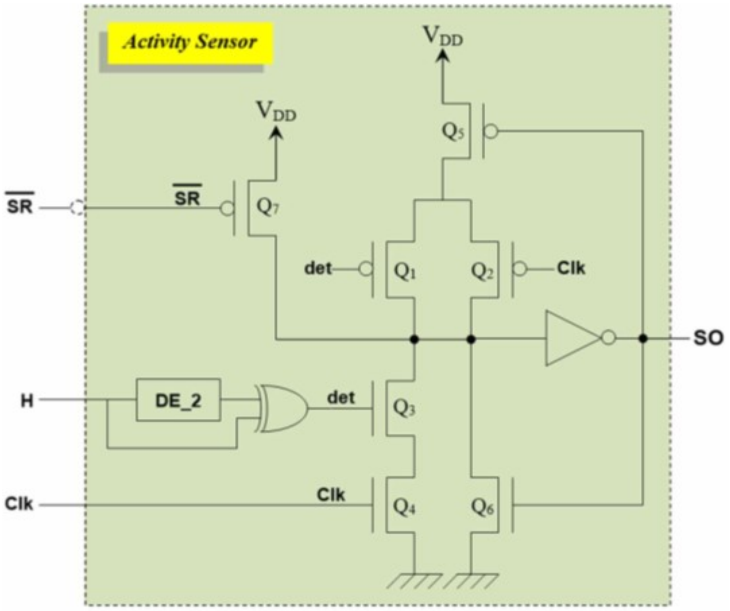


Figure 3.25: Activity sensor architecture.

3.4.2 Global Performance Sensor

The global performance sensor (GPS) for subthreshold operation, is constituted by two dummy critical paths (CP), in which a path is highly sensitive to NBTI degradations, while the other is highly sensitive to PBTI aging effects. By monitoring the delays in these two dummy critical paths, according with the available clock frequency and estimated PVTA degradation, it is possible evaluate the performance of the GPS under the working conditions and extrapolate for the main circuit. Moreover, by registering the correct output of the GPS evaluation of the dummy paths, it is also possible to know if the performance is relaxed for the available clock frequency and PVTA degradations, or if it is stressed and error occurrence is eminent. This sensor also uses an ultra-low-power AVS strategy, where the power supply voltage can be automatically adjusted during sensor operation.

Figure 3.26 shows the GPS architecture, which consists of a controller block, two dummy critical paths, and two groups of sensor latches.

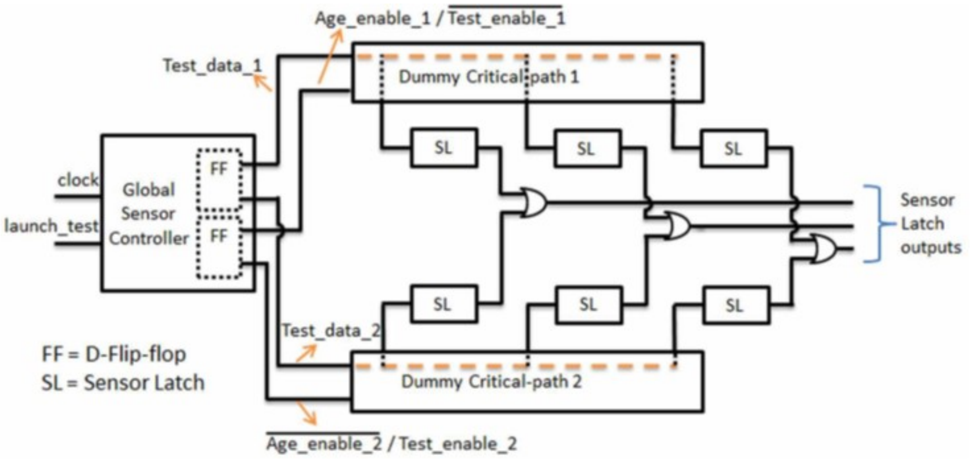


Figure 3.26: Global performance sensor architecture.

The controller block launches two consecutive signal transitions (Low-to-High and High-to-Low) in each dummy signal path, to trigger two different signal propagations in each dummy path. The sensor latches are placed and distributed along both dummy paths and will capture the signals along the paths for the available clock period. These sensor latches have activity sensor blocks (similar to LPS) to detect and signalize unsafe data captures in the latch. Therefore, the number of flagged sensor latches allows to evaluate the performance of the GPS dummy paths according with the available clock frequency and PVTA degradation. For a relaxed (low) clock frequency, no sensor latch will be signalized. However, for a stressed (high) clock frequency, the first sensor latches in the paths will be activated (detecting unsafe data captures). As we do not know which dummy critical path ages more, OR gates are used to connect the sensor latches from both dummy paths, to obtain one final sensor latch output and GPS output.

The two dummy critical paths can be seen in more detail in figure 3.27, where one of the dummy paths is implemented with NOR gates so that the NBTI aging degradation is greater and the other dummy path is implemented with NAND gates, creating a path with higher PBTI aging degradation.

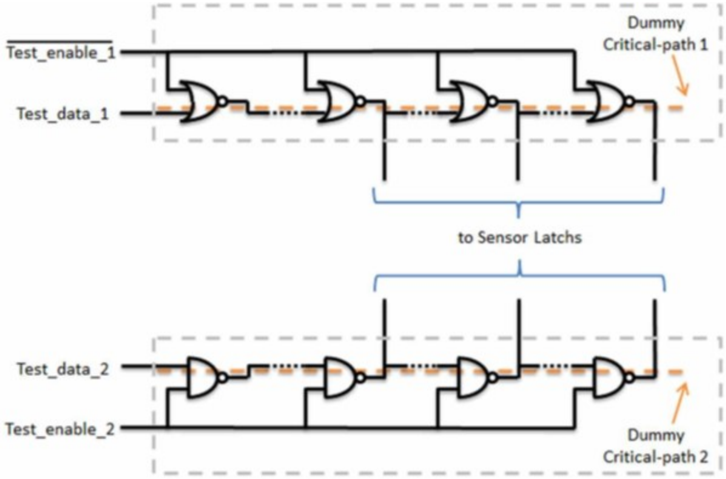


Figure 3.27: Detail of the dummy critical paths in the GPS.

3.4.3 Complete Performance Sensor Solution

In order to optimize energy efficiency, we can use GPS and LPS together to implement an adaptive voltage scaling (AVS) strategy, where, when GPS and LPS sense slowness and performance loss under high workload requirements, the controller acts by increasing supply voltage and when sensors sense high performance under low workload requirements, the controller acts by slowly decreasing supply voltage, until the edge of detection. This way the circuit supply voltage is regularly and permanently adjusted to its optimum value. The GPS can monitor regularly circuit operation and tune power-supply voltage accordingly, to obtain an efficient power consumption for the required circuit performance and workload. The LPS can also trigger power-supply voltage changes, but most importantly it monitors circuit performance locally, where functional errors may occur, and triggers GPS tuning.

Figure 3.28 summarizes a typical optimized circuit operation and how sensors (GPS and LPS) are used to keep the circuit running with the smallest V_{DD} value for each clock frequency.

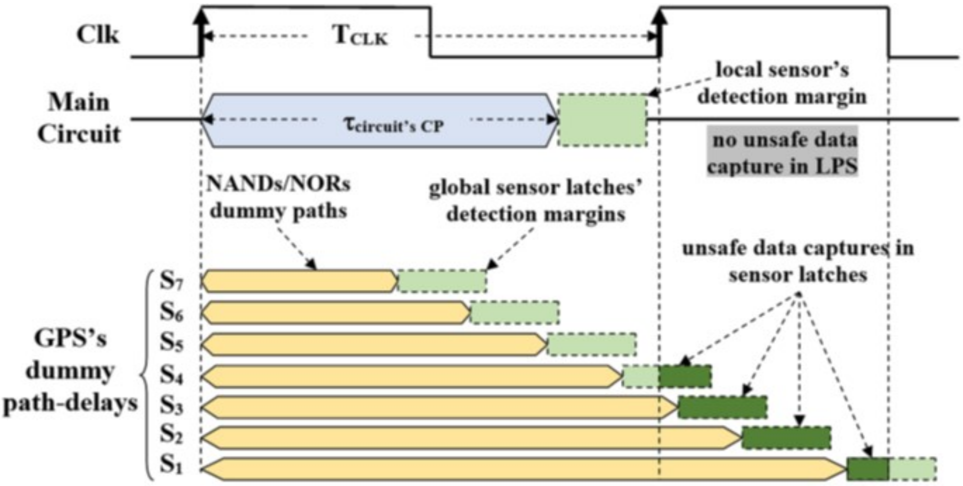

Clk
T_{CLK}
Main Circuit
$\tau_{circuit's\ CP}$
local sensor's detection margin
no unsafe data capture in LPS
NANDs/NORs dummy paths
global sensor latches' detection margins
unsafe data captures in sensor latches
GPS's dummy path-delays
S_7
S_6
S_5
S_4
S_3
S_2
S_1

Figure 3.28: Typical optimized circuit operation and sensor use.

The rising edge of the clock starts signal switching on the dummy paths and the critical path of the main circuit. These signal switches from a certain moment stop switching and become stable. There is a detection margin in GPS and LPS, for which the signals are not yet stables due to PVTA variations. When these detection margins reach the next rising edge of the clock (after a clock period), the sensor latches in GPS and the activity sensor in LPS captures unsafe data, signaling on output of the sensor the occurrence of an error. This is what happens in the sensor latches (S_1-S_4), because due to the decrease in the V_{DD}, the delay increased considerably implying an unstable signal capture on flip-flops. When we decrease the supply voltage (V_{DD}) it is necessary to decrease the frequency of the clock (increase of the clock period) so that no errors occur. This is our AVS strategy where the sensor is always comparing circuit delays with clock delays so that you can always work on the edge without the sensor failure.

Chapter 4

Study of Scout Memory Sensor for Subthreshold Voltages

Power consumption in CMOS integrated circuits, as never before, has a huge importance in today's chips for IoT applications, as all self-powered devices quest for the never-ending battery life, but also with smaller and smaller dimensions every day, in order to be used widely. Therefore, one of the objectives of this dissertation is the analysis of the behavior of memory cell sensors for subthreshold conditions, that is, for supply voltages close to the threshold voltage or lower.

In MOSFET model, it is assumed that current only flows through the MOSFET channel when $V_{GS} > V_T$. In reality, current flows even when V_{GS} is below the threshold voltage, but it is orders of magnitude weaker than currents in strong inversion. The inversion layer that is seen in strong inversion is barely seen in this case, and this regime can also be called weak inversion, where some electrons diffuse from the source into the channel (leakage current).

Under weak inversion, the relation between current and gate-source voltage becomes exponential (figure 4.1), according to the following equation:

$$I_D \propto exp\left(\frac{qV_{GS}}{nkT}\right) \tag{4.1}$$

Thus, it is possible that the sensors and the memory can still function under subthreshold.

In this chapter is made the study of the Scout Memory Sensor (presented in chapter 3) for supply voltages below the nominal voltage as is the case of the subthreshold regime. This chapter pretends to know if for lower voltages, the Scout Memory Sensor is still a robust solution, presenting reliable behavior and what the minimum supply voltage for which the sensor is reliable.

In this chapter the operation of each block of the Scout Memory Sensor is analyzed, namely the transition detector, the pulse detector and also the complete circuit of the sensor. Some simulations were made using Cadence software and UMC130 technology.

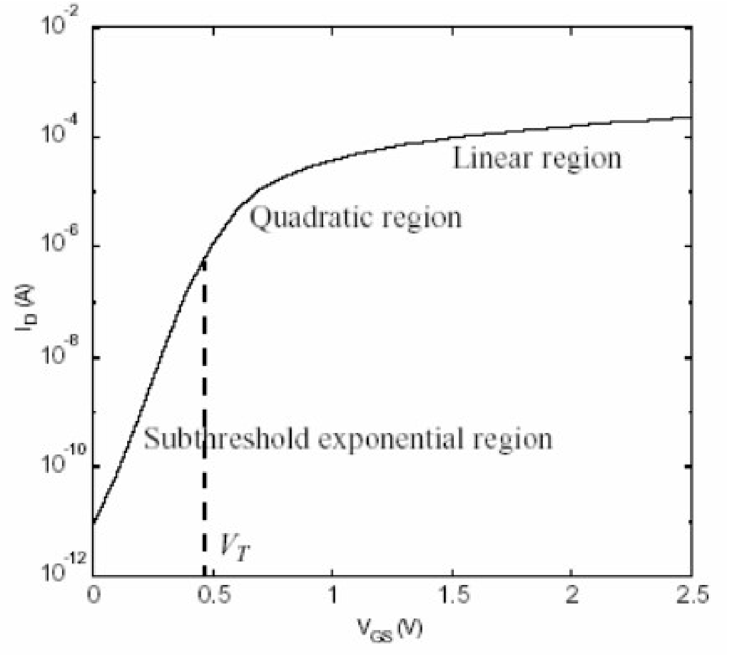


Figure 4.1: MOSFET I_D vs V_{GS} characteristic.

4.1 Transition Detector

The Scout Memory Sensor transition detector consists of a set of unbalanced n-type and p-type inverters. Note that a p-type inverter is an inverter with a more conductive PMOS when compared with NMOS, while an n-type inverter is the reverse situation, i.e., it contains a more conductive NMOS when compared with PMOS. In the connecting nodes of these inverters there are parasitic capacities always with the same value, which are connected to power supply (V_{DD}) by PMOS transistors (figure 4.2). When you lower the V_{DD} value it is natural for the circuit to become slower, because as the supplied energy is lower the current will also be lower and the parasitic capacities take longer to load or discharge. Thus, it is to be expected that the pulses generated by the transition detector will have more delay as the supply voltage decreases.

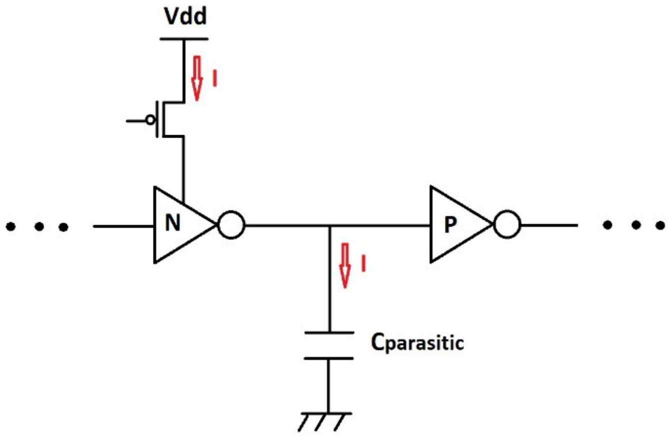


Figure 4.2: Connection nodes of the transition detector inverters.

On the other hand, for a transition from the bit line from ”Low” to ”High” as the top path of the inverters begins with an n-type, this means that the output signal of the inverter switches earlier, while for the bottom path of the inverters, the output signal of a p-type inverter switches later (figure 4.3). The sum of all Δt caused by the unbalanced inverters is proportional to the pulse width at the output XOR and when the supply voltage (V_{DD}) decreases, all internal switches are slower, therefore the Δt of each pair of inverters will be higher and consequently we will get a pulse with greater width at the output of the transition detector.

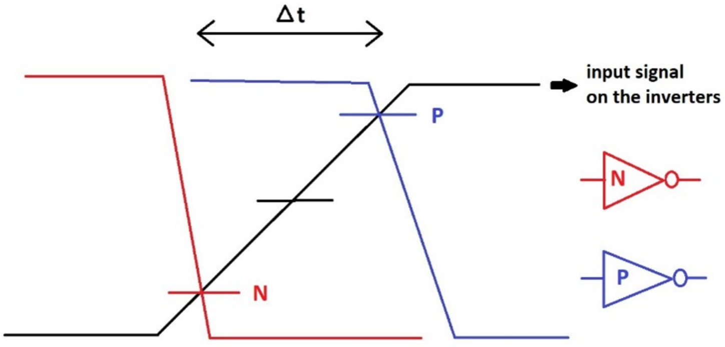
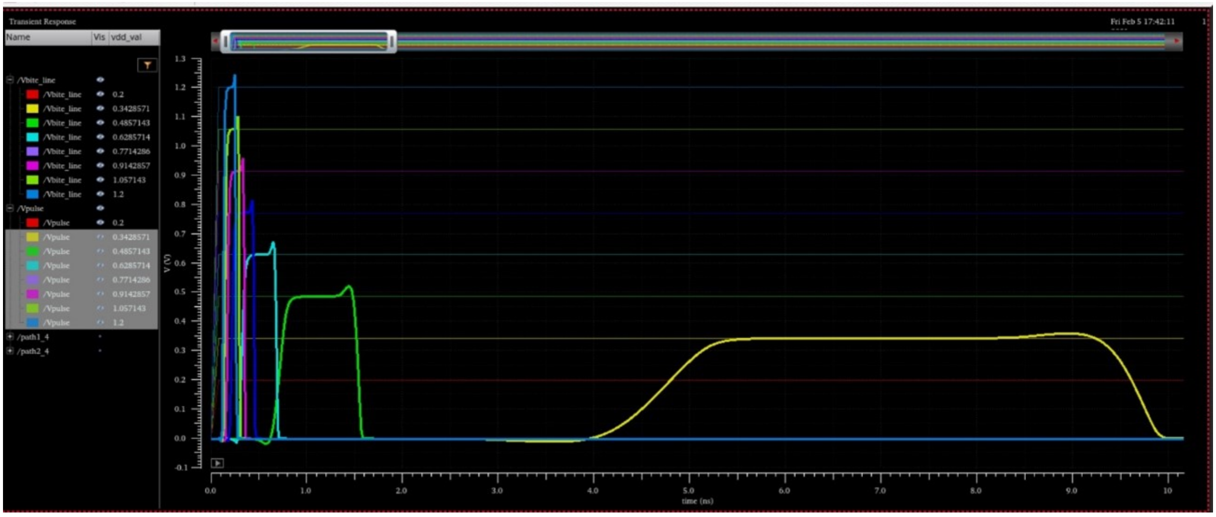


Figure 4.3: Internal switches of the transition detector inverters.

4.1.1 Simulation Results

In order to study the operation of the transition detector for supply voltages below the nominal voltage, a parametric simulation was carried out for a transition from the bit line from ”Low” to ”High” with a rise time of 80 ps, in which the following sizes were used for the inverters: WNmin=160nm, WPmin=600nm. The graphics obtained for the output of the transition detector are shown below and refer to supply voltages with $V_{DD} \in [0,34V;1,2V]$ and also with $V_{DD} \in [0,1V;0,2V]$ for a better detail in the situation of subthreshold voltages.


Transient Response

Name	Vis	vdd_val
/Vbite_line		
🟥 /Vbite_line		0.2
🟨 /Vbite_line		0.3428571
🟩 /Vbite_line		0.4857143
/Vbite_line		0.6285714
/Vbite_line		0.7714286
/Vbite_line		0.9142857
/Vbite_line		1.057143
🟦 /Vbite_line		1.2
/Vpulse		
🟥 /Vpulse		0.2
🟨 /Vpulse		0.3428571
🟩 /Vpulse		0.4857143
/Vpulse		0.6285714
/Vpulse		0.7714286
/Vpulse		0.9142857
/Vpulse		1.057143
🟦 /Vpulse		1.2
/path1_4		
/path2_4		

Figure 4.4: Parametric analysis of the transition detector for $V_{DD} \in [0,34V;1,2V]$.

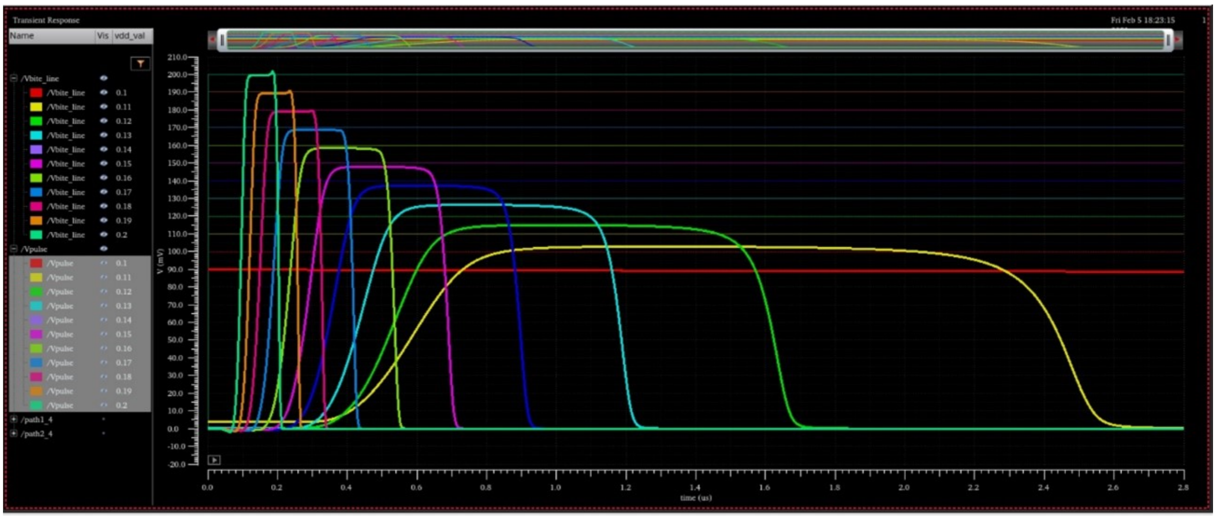


Figure 4.5: Parametric analysis of the transition detector for $V_{DD} \in [0,1V;0,2V]$.

Vdd	Pulse delay	Pulse width
1,2 V	102,78 ps	130,28 ps
1,06 V	121,38 ps	146,94 ps
0,91 V	145,41 ps	175,33 ps
0,77 V	185,82 ps	231,24 ps
0,63 V	272,05 ps	358,58 ps
0,49 V	621,81 ps	811,71 ps
0,34 V	3,98 ns	4,98 ns

Table 4.1: Parametric analysis data of the transition detector for $V_{DD} \in [0,34V;1,2V]$.

Through the graphics it is possible to observe that by decreasing the supply voltage, the pulse generated by the transition detector (Vpulse) has greater delay and its width gradually increases for the reasons mentioned above. Figure 4.5 shows that the transition detector still works for subthreshold voltages, but for V_{DD} values below 0.1 V the transition detector stops generating a pulse, making it impossible to function in this range of values. For better analysis of the graphics, tables 4.1 and 4.2 are presented with data related to the pulses generated by the transition detector, for various V_{DD} values.

Vdd	Pulse delay	Pulse width
0,2 V	72,24 ns	107,23 ns
0,19 V	85,74 ns	136,87 ns
0,18 V	107,76 ns	176,55 ns
0,17 V	130,32 ns	228,75 ns
0,16 V	163,52 ns	298,69 ns
0,15 V	195,89 ns	395,11 ns
0,14 V	241,61 ns	528,54 ns
0,13 V	298,26 ns	736,12 ns
0,12 V	346,11 ns	1,08 μs
0,11 V	395,49 ns	1,84 μs
0,1 V	—	—

Table 4.2: Parametric analysis data of the transition detector for $V_{DD} \in [0,1V;0,2V]$.

4.2 Pulse Detector

The pulse detector converts the pulse generated by the transition detector (Vpulse) into a DC voltage (Vsense) by charging a capacitor (C1) by a PMOS transistor. When the V_{DD} value decreases, we will obtain pulses with longer duration and consequently the C1 capacitor will be charged with different growth rates, which implies that different values of Vsense will be obtained.

4.2.1 Simulation Results

To analyze the pulse detector behavior for various V_{DD} values, a parametric simulation of the transition detector together with the pulse detector was performed, for a bit line signal from ”Low” to ”High” with a rise time of 80 ps. Transistors with a size of WNmin=160nm, WPmin=600nm and a capacitor with a capacity of 17 fF were used. In this simulation, the maximum sensitivity of the sensor was used, i.e., the three control bits were activated. The simulation result is shown in the graphic below for $V_{DD} \in [0,34V;1,2V]$ values.

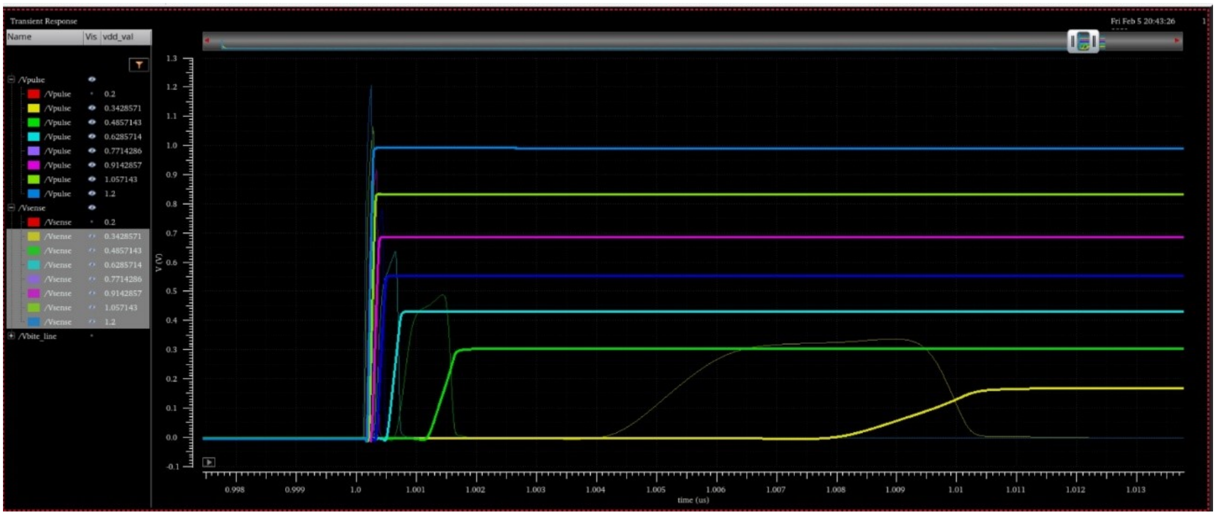


Figure 4.6: Parametric analysis of the pulse detector for $V_{DD} \in [0,34V;1,2V]$.

By observing the graphic, it is possible to note that the value of Vsense has a higher delay for smaller supply voltages, due to the fact that less power is provided, all parasitic capacities take longer to charge and discharge. On the other hand, the value of Vsense decreases to smaller supply voltages and its growth rate also decreases. To analyze the graphic data in more detail, a table is displayed with the numerical values of the Vsense signal and its percentage in relation to the V_{DD}.

The data in table 4.3 confirm that Vsense decreases to lower supply voltages and that the percentage of Vsense relative to V_{DD} also decreases, which means that the capacitor C1 takes longer to charge for smaller supply voltages even though the duration of the pulses (Vpulse) is longer. This pulse detector block still works for subthreshold voltages, as long as the pulse (Vpulse) is generated in the transition detector, a DC signal to the output is produced but with greater delay. Although the pulse detector continues to work for subthreshold voltages, as the percentage of Vsense relative to V_{DD} has decreased,

Vdd	Vsense	Vsense in % of Vdd
1,2 V	990 mV	82,5 %
1,06 V	833 mV	78,6 %
0,91 V	687 mV	75,5 %
0,77 V	555 mV	72,1 %
0,63 V	433 mV	68,7 %
0,49 V	304 mV	62,1 %
0,34 V	170 mV	50 %
0,2 V	103 mV	51,5 %

Table 4.3: Parametric analysis data of the pulse detector for $V_{DD} \in [0,2V;1,2V]$.

the DC voltage generated at the output of the pulse detector becomes extremely low in the subthreshold regime.

4.3 Complete Circuit

In this section, the complete circuit of the Scout Memory Sensor is analyzed for voltages below the nominal voltage. When we lower the supply voltage to values below the subthreshold voltage it is natural that the value of Vsense and Vref reach extremely low values, which can lead to unsatisfactory results when the comparator is activated.

4.3.1 Simulation Results

To study the sensor behavior for low supply voltage, a parametric simulation of the 4 sensor blocks was performed for a bit line of ”Low” to ”High” with a rise time of 700 ps. Capacitors with capacities C1=17 fF, C2=100 fF, transistors with size WNmin=160 nm, WPmin=600 nm and the three control bits activated were used.

The results are displayed in a graphic and in a table for $V_{DD} \in [0,2V;1,2V]$ values.

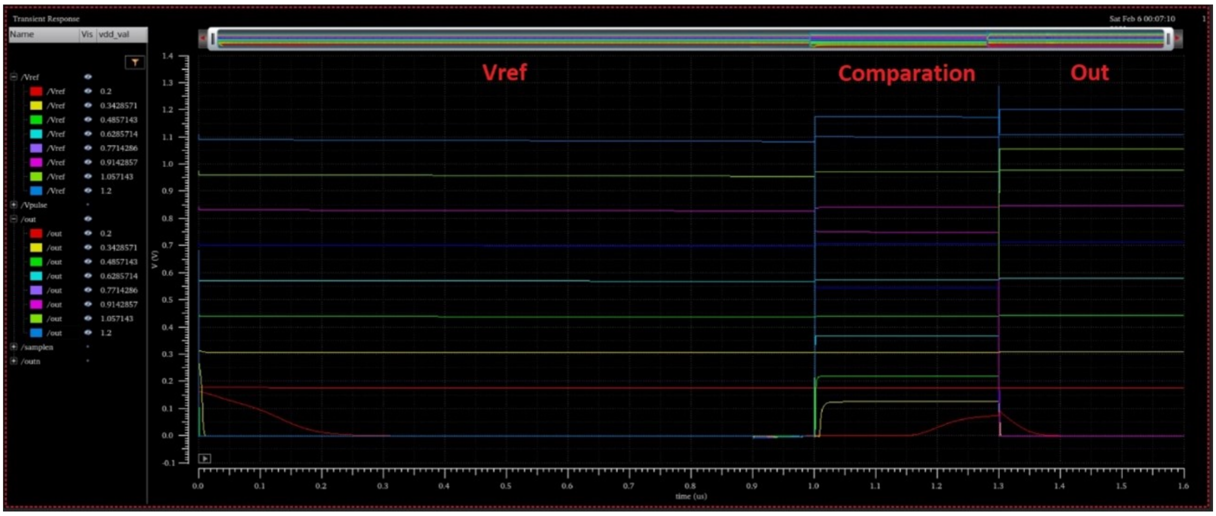


Figure 4.7: Parametric analysis of the complete circuit for $V_{DD} \in [0,2V;1,2V]$.

Vdd	Pulse Delay	Pulse Width	Vsense	Vref	Out	Vsense in % of Vdd	Vref in % of Vdd
1,2 V	388,99 ps	248,55 ps	1,17 V	1,09 V	1,2 V	97,94 %	91,13 %
1,06 V	427,81 ps	238,37 ps	971 mV	965 mV	1,06 V	91,6 %	91,04 %
0,91 V	478,04 ps	239,71 ps	749 m V	828 mV	8,66 μV	82,31 %	90,99 %
0,77 V	546,32 ps	267,45 ps	545 mV	699 mV	3,82 μV	70,78 %	90,78 %
0,63 V	708,81 ps	357,28 ps	368 mV	569 mV	4,14 μV	58,41 %	90,32 %
0,49 V	1,11 ns	725,07 ps	220 mV	438 mV	6,52 μV	44,9 %	89,39 %
0,34 V	4,46 ns	4,55 ns	127 mV	302 mV	4,14 μV	37,35 %	88,82 %
0,2 V	77,68 ns	98,09 ns	75 mV	176 mV	6,52 μV	37,54 %	88 %

Table 4.4: Parametric analysis data of the complete circuit for $V_{DD} \in [0,2V; 1,2V]$.

Looking at the graphic and data in table 4.4, we can verify that for a slow transition of the bit line (700 ps), the sensor only signals error for V_{DD}=1.2 V and V_{DD}=1.06 V. We can therefore conclude that this sensor does not work correctly when the supply voltage drops, because when the sensor detects an error, it should signal it to the full range of V_{DD} values. It is therefore necessary to modify the structure of the Scout Memory Sensor, namely the comparator block where failures happen for lower supply voltages, due to the complexity of its architecture. Another observation we can draw from the data is that the percentage of Vref relative to V_{DD} decreases as you would expect, but in the subthreshold region the Vref becomes approximately zero, making the operation of the sensor virtually impossible.

4.4 Scout Memory Sensor Analysis

Through the simulations carried out we can conclude that the transition detector and pulse detector blocks still work under subthreshold, and the minimum supply voltage for which the sensor still works is 0.11 V. In the subthreshold regime, the transition detector can still generate a pulse, albeit with a high delay. But simulations also show that the comparator block presents problems for subthreshold voltages due to the complexity of its architecture and the stack of electronic components. The Scout Memory Sensor is no longer consistent and coherent in the subthreshold regime, as it does not guarantee error signaling when supply voltages are very low for too slow bit line transitions. In addition, for very low supply voltages the difference between the Vsense and Vref signals becomes too small for the comparator block to function properly.

For these reasons, the Scout Memory Sensor is an incomplete sensor and fails under certain conditions. Therefore, it is necessary to find an alternative to this sensor that works at lower supply voltages. This dissertation aims to present new solutions to the problems verified in the Scout Memory Sensor and we can also take advantage of the transition detector block, because this block works correctly under subthreshold voltage values.

Chapter 5

Ultra-Low Power Performance Sensor for Memories

In this chapter a new sensor is presented to overcome the problems detected in the Scout Memory Sensor (chapter 4). This sensor is compatible with various types of memory and architectures (SRAM and DRAM) and is a performance sensor that detects the degradation caused by PVTA variations with low power consumption, and is compatible with DVFS (Dynamic Voltage and Frequency Scaling) techniques, thus allowing its use for smaller supply voltages (V_{DD}) in order to save energy. This Ultra-Low Power Performance Sensor is a novelty compared to the previously proposed sensors, so it has not yet been tested on real circuits.

The basic idea of this sensor is to make use of the clock synchronism that starts the operations of writing or reading the memory, to have a reference for measuring performance for the available V_{DD}. Hence, in this work we consider that first rising edge of the clock will trigger the signals to read/write in the memory (R/W), as depicted on figure 5.1. These signals will allow the writing/reading of the memory, which will cause a transition on the bit line (BL). Moreover, this bit line data can be used by flip-flops or other combinational logic to perform whatever is needed in the circuit function (use of BL data). All these processes and data reading/writing of signals add propagation delays to the initial instant of the first rising edge of the clock, corresponding to the normal operation of the circuit, and consequently corresponds to the critical path of the circuit. This normal operation must be faster than a clock period (T_{CLK}), so that an error does not occur during the read/write operation of the memory. Therefore, an additional safety margin is added to the critical path, to define the clock period (T_{CLK}), to avoid errors and account for any unpredictable PVTA variation. Typically, a PVTA variation causes all delays to change, and when this safety margin is reduced to zero, it means that an error will occur, because the normal functionality does not have time to be completed during the clock period. The sensor needs to work predictively, by working in this safety margin and detect the reduction of this margin to unsafe levels, i.e., to a minimum allowed safety margin. When this minimum safety margin is reached, the sensor signalizes an error, indicating that corrective measures (like V_{DD} increase, or frequency decrease) should be taken, otherwise this predictive error (only detected by the sensor) is transformed in an actual error (affecting the normal operation of the circuit).

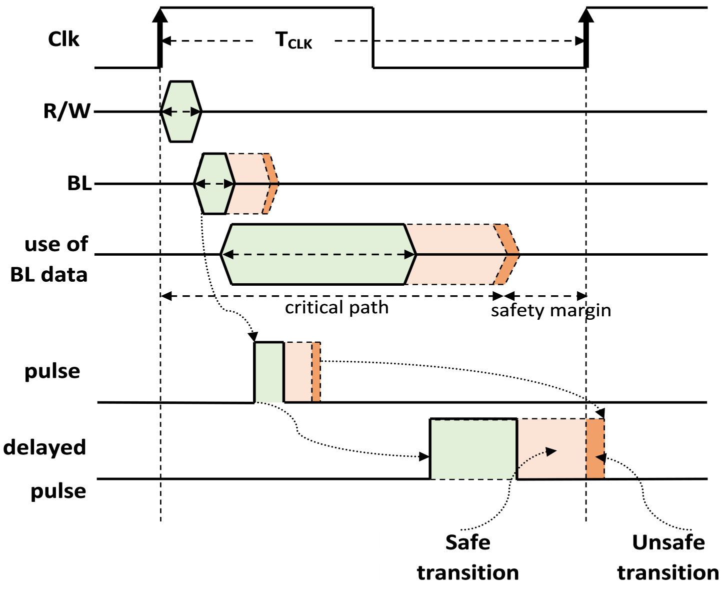


Figure 5.1: Map of delays in the Ultra-Low Power Performance Sensor within the clock period.

The described behavior of the sensor, shows that it is appropriate to be used in a DVFS technique, not only because it works at reduced V_{DD} voltage values, but also because it allows to control online the appropriate V_{DD} for each clock frequency used, reducing power to the minimum required by controlling the minimum safety margin used. In fact, there are two variables that the designer/user can control: V_{DD} and clock frequency. When we reduce V_{DD}, all the delays increase and so it is necessary to decrease the frequency of the clock, for the sensor to work at maximum performance, i.e. for each V_{DD} value there is a clock frequency from which the sensor is on the imminence of indicating an error, which for a small variation of PVTA conditions, the sensor may indicate an error. The purpose of this sensor is to detect an error, for a given V_{DD} and a certain frequency, always working at the limit with a small minimum safety margin. In this work, we consider that for a correct operation of a memory with the sensor, no error output should be produced, so an additional safety margin of, approximately, 20% of the clock period, was always used in a normal operation. Therefore, there should be optimal pairs of V_{DD}/f_{CLK} to work at the eminence of an error detection (and with this minimum safety margin), and the control strategy used should impose changes in V_{DD} or in f_{CLK} according to the optimization purpose and existing PVTA variation, allowing to: work at reduced power, or work at best performance.

5.1 Sensor Architecture

The architecture of the Ultra-Low Power Performance Sensor is composed of: a transition detector block, a delay element block, and a D flip-flop. Figure 5.2 presents these basic blocks that make up the Ultra-Low Power Performance Sensor, and we can see that it is a simpler architecture, when compared with the previous Scout Memory Sensor. As mentioned, we consider that the clock signal triggers all the operations in the memory cell (read/write), and therefore it triggers all the signal transitions that occur in the sensor. When a bit line changes its value due to a read/write operation, the transition detector block generates a pulse with a width proportional to the bit line transition delay; then, a delay element block delays the generated pulse with a delay that makes the delayed pulse to reach the safety margin window, considering that the circuit is working in the eminence of an error, as explained in Figure 5.1; if the flip-flop captures the pulse, then an unsafe transition is signalized.

Note that the ending of the delayed pulse should always occur after the critical path of the circuit makes his last change before the new clock trigger, and this should be defined by design. Moreover, in this work it was not consider the possibility of online changing sensor's sensibility. However, this is possible to implement and it will be addressed in future work.

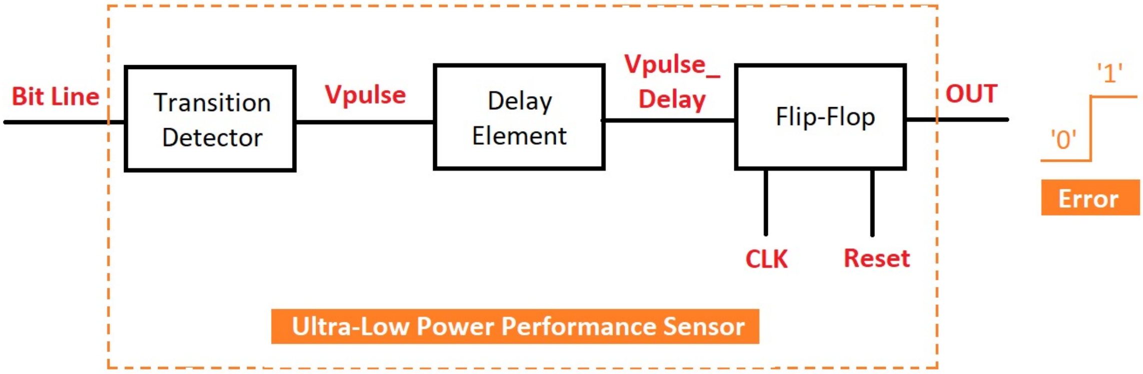


Figure 5.2: Ultra-Low Power Performance Sensor architecture.

5.1.1 Transition Detector

The transition detector block is the same block used in the Scout Memory Sensor, because this block performed well for low V_{DD} values. However, according with the type of memory where the sensor is used, and on the sensor implementation strategy (if all the memory cells are monitored, or if only a random sample are monitored), the transition detector can have three different architectures:

- Version 1: transition detector to monitor 1 bit line's SRAM cell initialized to V_{DD};
- Version 2: transition detector to monitor 2 bit lines' SRAM cell initialized to V_{DD};
- Version 3: transition detector to monitor 1 bit line's DRAM cell initialized to V_{DD}/2.

Other versions could also be developed, but we think these 3 versions cover the main cases of application of the sensor in SRAM and DRAM memory cells.

Version 1: Transition Detector to Monitor 1 Bit Line's SRAM Cell Initialized to V_{DD}

The first implementation of the transition detector is used for SRAM memories initialized to V_{DD} that only monitors transitions in one of the two bit lines from a cell. Although an SRAM cell has two bit lines, this implementation is considered for the case when not all the cells in a memory are monitored, but a random sample are chosen to be monitored, and the results are given for a statistic representation of the memory.

The problem of monitoring only one bit line of an SRAM memory cell is the fact that the bit-lines are initialized at V_{DD}, and in a reading/writing operation only the reading/writing of V_{SS} (a logic 0) may produce a transition in the bit line, and consequently generate a pulse in the transition detector. In other words, the reading/writing of a logic 1 (V_{DD}) does not activates the sensor and the monitoring procedure is masked. The solution to overcome this problem is to monitor both bit lines of the cell, because the complementary bit line will read/write a logic 0, activating the sensor. But, this is the version 2 implementation, that will be discussed later on.

Nevertheless, considering that a random number of bit lines are chosen to be monitored, the selected bit lines should statistically represent the memory, so that the available sensors can predictively detect unsafe transitions and this information can represent all the cells in the memory. In this case, it is wiser to use a higher safety margin in the circuit operation.

The structure of this first implementation is shown in figure 5.3.

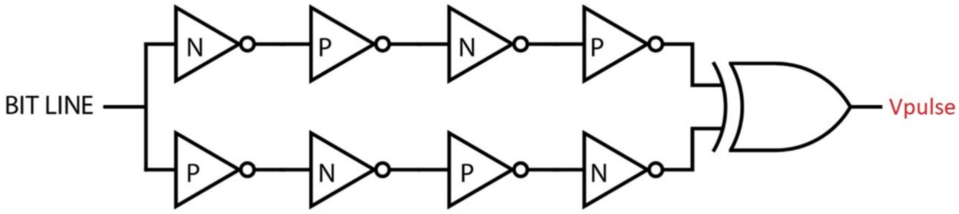


Figure 5.3: Transition detector to monitor 1 bit line's SRAM cell initialized to V_{DD}.

As described in previous chapters, the transition detector generates a pulse, with a width proportional to the transition time of the bit line. In figure 5.3, the bottom and top path has 4 inverters, but a higher number can be used to increase the width of the generated pulses, so that the sensor more easily detects the imminence of an error. Changing the number of inverters in the chains will change the sensitivity of the sensor, as the pulses will also change their width.

To illustrate the operation of the transition detector, a parametric simulation was carried out for various V_{DD} values, using a writing operation of the value ”0”, in a SRAM memory cell (figure 5.4 and 5.5). Transistors with size WNmin=160nm and WPmin=600nm were used and the SRAM memory was initialized to V_{DD} with the help of the Pre-Charge circuit.

Figure 5.4 shows that for lower V_{DD} values, the pulse generated by the transition detector has a higher delay and a greater width. This is a good and important result, because it indicates that the sensor gets more sensitive when the working conditions gets worse. In other words, there is not a unique minimum safety margin for all the working conditions, and the minimum safety margin increases when the conditions are worse (in this case, when V_{DD} decreases).

In Figure 5.4 it is also possible to observe that for a V_{DD} of 0.4V, it is still possible to detect a bit line

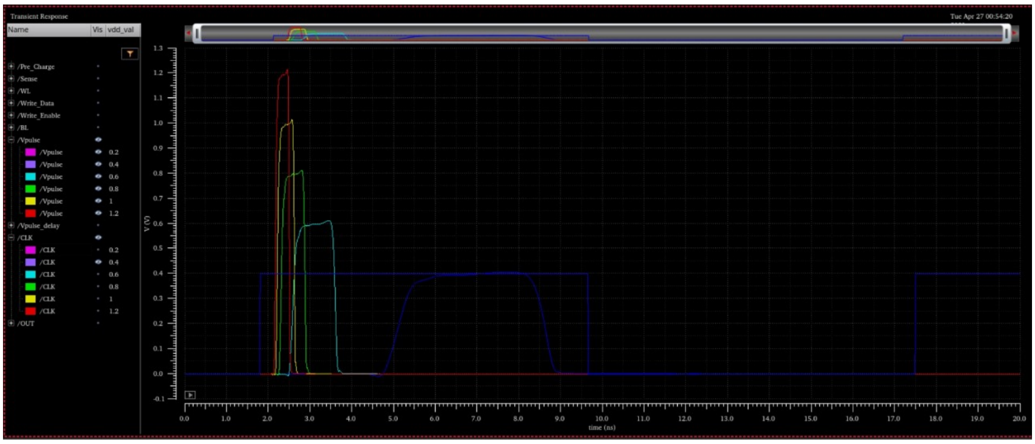


Figure 5.4: Parametric analysis of transition detector in SRAM memory for $V_{DD} \in [0,4V;1,2V]$.

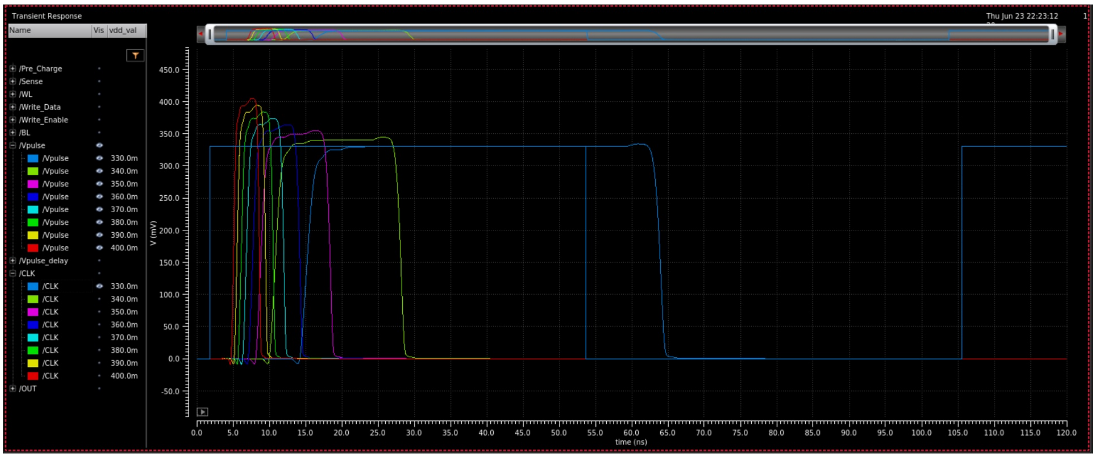


Figure 5.5: Parametric analysis of transition detector in SRAM memory for $V_{DD} \in [0,33V;0,4V]$.

transition. Through figure 5.5, it is also possible to conclude that the minimum supply voltage for the transistor detector to work is 0.33V, because below this value, there are no more transitions in the bit line and therefore it is not possible to generate a pulse in the transition detector.

Version 2: Transition Detector to Monitor 2 Bit Lines' SRAM Cell Initialized to V_{DD}

The second implementation of the transition detector can be used for SRAM memories initialized to V_{DD} to monitor both cell's bit lines. This implementation allows to monitor memory cells in a complete way, i.e, for all bits stored in the memory. As it was mentioned before, because SRAM cells' bit lines are initialized at V_{DD}, when a read/write operation occurs, the bit lines will only change if a logic 0 (V_{SS}) is read/wrote. However, if one bit line is at logic 1 (V_{DD}), the complemented bit line is at logic 0 (V_{SS}). Therefore, if both bit lines of a memory cell are monitored, the sensor can always be stimulated in a read/write operation. One can argue that if we consider circuit aging, not all transistors are tested by the sensor. However, we can also argue back that if the transistors are not used in the reading/writing operations, they will not be used to produce an error in the reading/writing operation, so the sensor monitors the transistors that influence cell behavior.

This version 2 implementation, though, has the disadvantage of increasing sensor circuit area and complexity, because both bit lines need to be monitored. The architecture of this implementation is shown in figure 5.6.

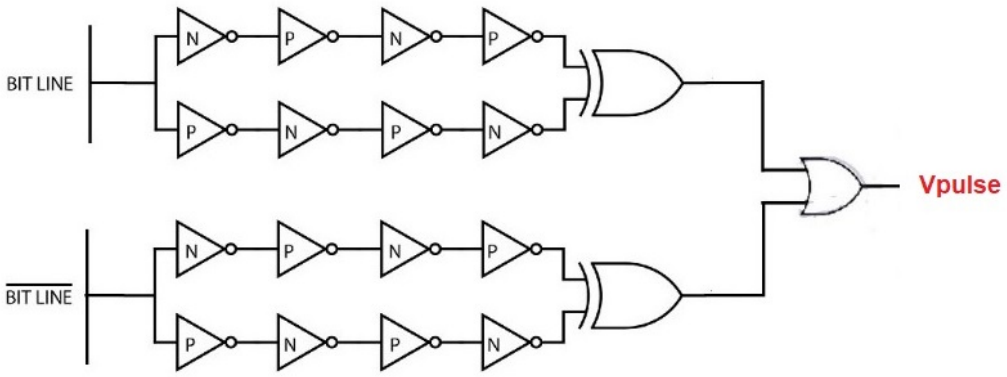


Figure 5.6: Transition detector to monitor 2 bit lines' SRAM cell initialized to V_{DD}.

Figure 5.6 shows that this new implementation uses two version 1 transition detectors, each connected to $bit\ line$ and the $\overline{bit\ line}$, in order to monitor both bit lines. Finally, an OR gate is used to generate Vpulse.

The simulation results of this implementation are very similar to the version 1 implementation, and no noticeable changes are observed, because the SRAM only activates one bit line at a time, and the propagation delay of the OR gate is the only slightly difference.

Version 3: transition detector to monitor 1 bit line's DRAM cell initialized to V_{DD}/2

The third implementation is used for DRAM memory cells, which are initialized to V_{DD}/2. As there is only one bit line connected to the memory cell, only a simple transition detector is required, but it needs to be connected to an AND gate, with its second input connected to the sense amplifier activation signal. This additional AND gate is to activate the transition detector only when the bit line changes with the read/write data, and not when the bit lines are initialized at V_{DD}/2. Note that because of the use of two different unbalanced inverters connected to the bit line (the N-type and the P-type), with a stable V_{DD}/2 value in the bit line, it would produce a stable High output at Vpulse, not creating the pulse proportional to the transition time. Thus, with this additional AND gate, the transition detector only generates a pulse when the sense amplifier is activated, so it does not generate problems in the initialization of the bit line to V_{DD}/2. The structure for this implementation can be seen through Figure 5.7.

In order to analyze the behavior of the sensor in a DRAM memory, a parametric simulation was carried out for various V_{DD} values, using a writing operation of the value ”0”, in a DRAM memory cell (figure 5.8). Transistors with size WNmin=160nm and WPmin=600nm were used and the DRAM memory was initialized to V_{DD}/2 with the help of the Pre-Charge circuit.

Looking at figure 5.8 we can verify that the transition detector continues to work normally, generating

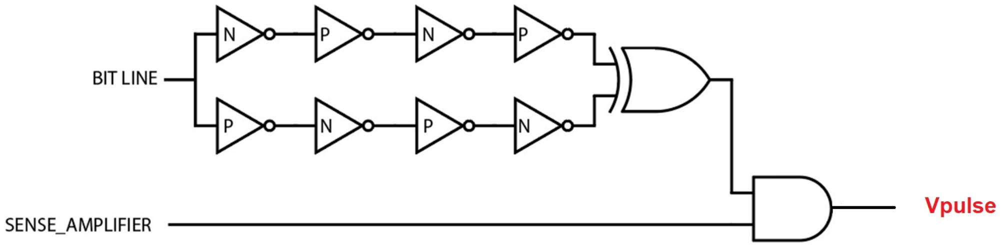


Figure 5.7: Transition detector to monitor 1 bit line's DRAM cell initialized to V_{DD}/2.

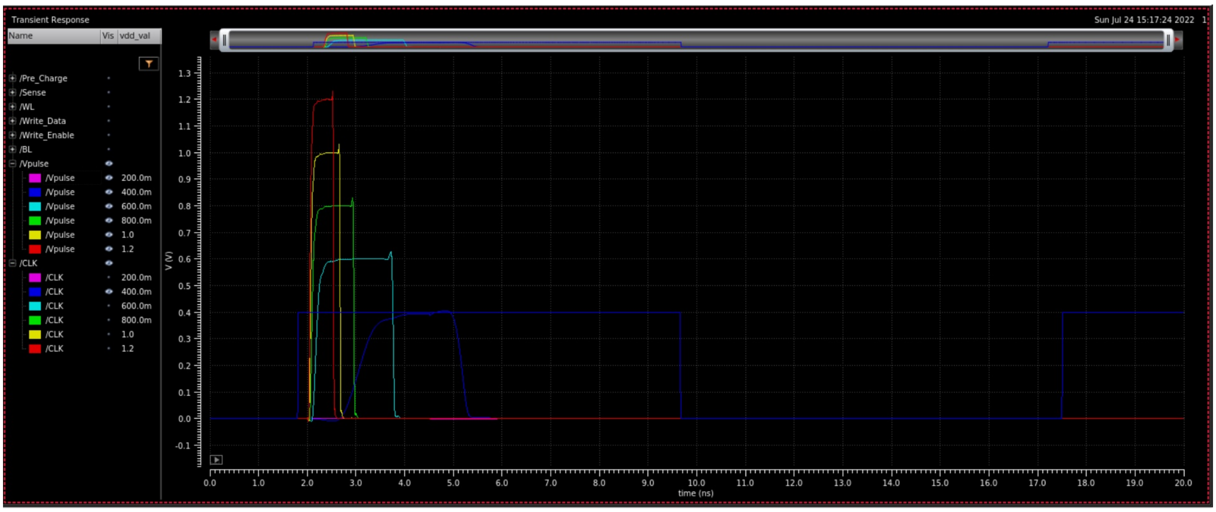


Figure 5.8: Parametric analysis of transition detector in DRAM memory for $V_{DD} \in [0,4V;1,2V]$ (writing operation of the value "0").

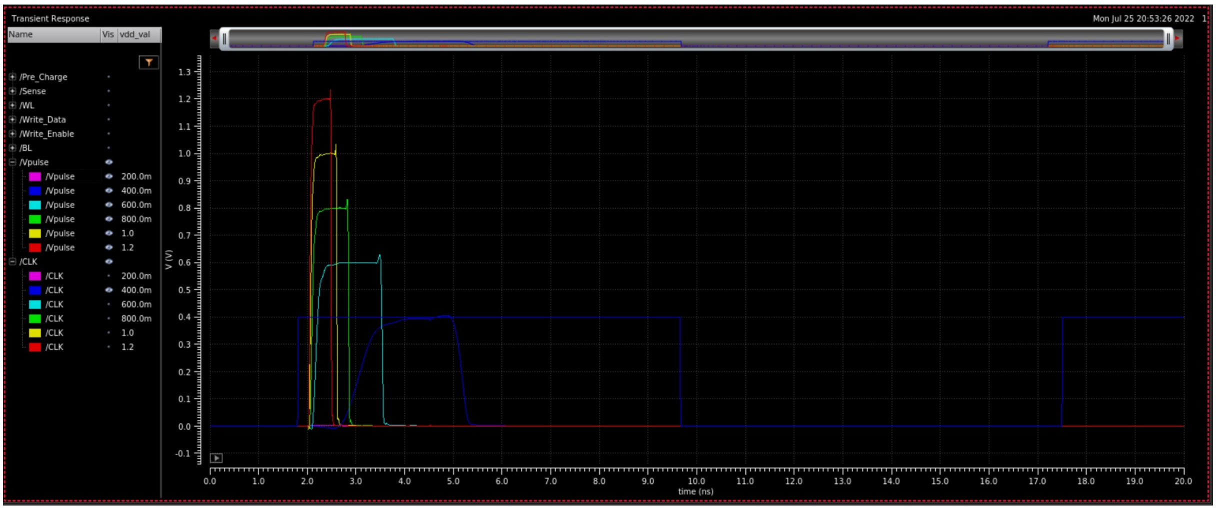
pulses for each bit line transition, but compared to figure 5.4, this time the pulses have lower width and a lower delay. This is because, as the DRAM memory is initialized to V_{DD}/2, the transition detector has to generate a pulse for a half transition. However, if needed, we can adjust the sensitivity of the transition detector by using more unbalanced inverters in the chains, allowing to obtain larger pulses.

DRAM memories have an advantage over SRAM memories, because to detect transitions from V_{SS} to V_{DD} and from V_{DD} to V_{SS}, the SRAM sensor must be connected to both bit lines for a SRAM memory. But in the case of a DRAM memory, the sensor only needs to be connected to a bit line, i.e., no need to add auxiliary hardware, which implies a smaller area overhead.

To check the operation of the transition detector for transitions of the bit line from "V_{DD}/2" to "V_{DD}", a parametric simulation was carried out for various V_{DD} values, but this time using a writing operation of the value "1", in a DRAM memory cell (figure 5.9). Transistors with size WNmin=160nm and WPmin=600nm were used and the DRAM memory was initialized to V_{DD}/2 with the help of the Pre-Charge circuit.

Comparing figures 5.9 and 5.8, it is possible to observe that the pulses are virtually identical, with the same width and delay, so the transition detector works equally for writings of "0" or "1".

As we can see, just by changing the implementation of this first block of the performance sensor, the transition detector block, we can adapt the usability of the sensor to different memories and different initialization values of the bit line. Moreover, all the versions of the transition detector block show similar



Transient Response

Mon Jul 25 20:53:26 2022 1

Name	Vis	vdd_val
/Pre_Charge	·	
/Sense	·	
/WL	·	
/Write_Data	·	
/Write_Enable	·	
/BL	·	
/Vpulse	◉	
🟪 /Vpulse	◉	200.0m
🟦 /Vpulse	◉	400.0m
▨ /Vpulse	◉	600.0m
🟩 /Vpulse	◉	800.0m
🟨 /Vpulse	◉	1.0
🟥 /Vpulse	◉	1.2
/CLK	◉	
🟪 /CLK	·	200.0m
🟦 /CLK	◉	400.0m
▨ /CLK	·	600.0m
🟩 /CLK	·	800.0m
🟨 /CLK	·	1.0
🟥 /CLK	·	1.2

Figure 5.9: Parametric analysis of transition detector in DRAM memory for $V_{DD} \in [0,4V;1,2V]$ (writing operation of the value "1").

output pulses, whether we use it in an SRAM or a DRAM memory cell. Therefore, for simplicity, in the next sections and when the reminding blocks of the sensor are presented, only the SRAM with version 1 transition detector simulation examples are presented, because the other versions will produce similar results for the complete sensor.

5.1.2 Delay Element

Reusing the same solution from previous works, this new performance sensor contains two delay elements of type H (DE_H), as presented in [19]. Each delay element consists of two inverters, connected to two transmission gates, as can be seen in figure 5.10.

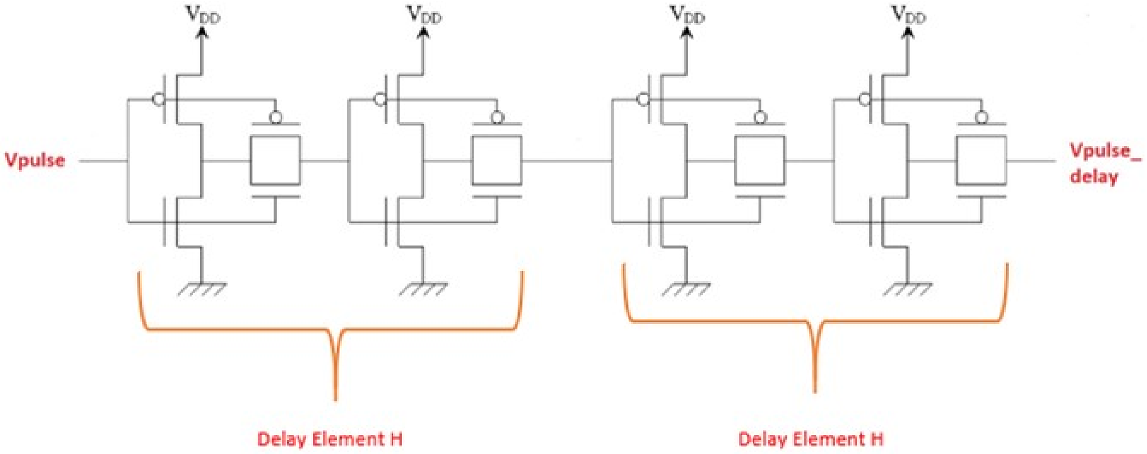


Figure 5.10: Two delay elements for Ultra-Low Power Performance Sensor.

The introduction of these two delay elements aims to create a delay in the pulses generated by the transition detector, so that there is sufficient time, for which the data stored in the memory cells, during read/write operations, can be used by flip-flops or other combinational logic of the circuit (the critical path). Note that, according with the critical path delay of the circuit, additional delay elements can be used, to introduce a higher delay in the pulse. As already explained in this section, the delayed pulse should be placed in the safety margin of the circuit, so that it can signalize predictively any error.

To show the output of the delay element block, a parametric simulation was performed for various V_{DD} values, using a writing operation of the value ”0”, in a SRAM memory cell (figure 5.11). Transistors with size WNmin=160nm and WPmin=600nm were used.

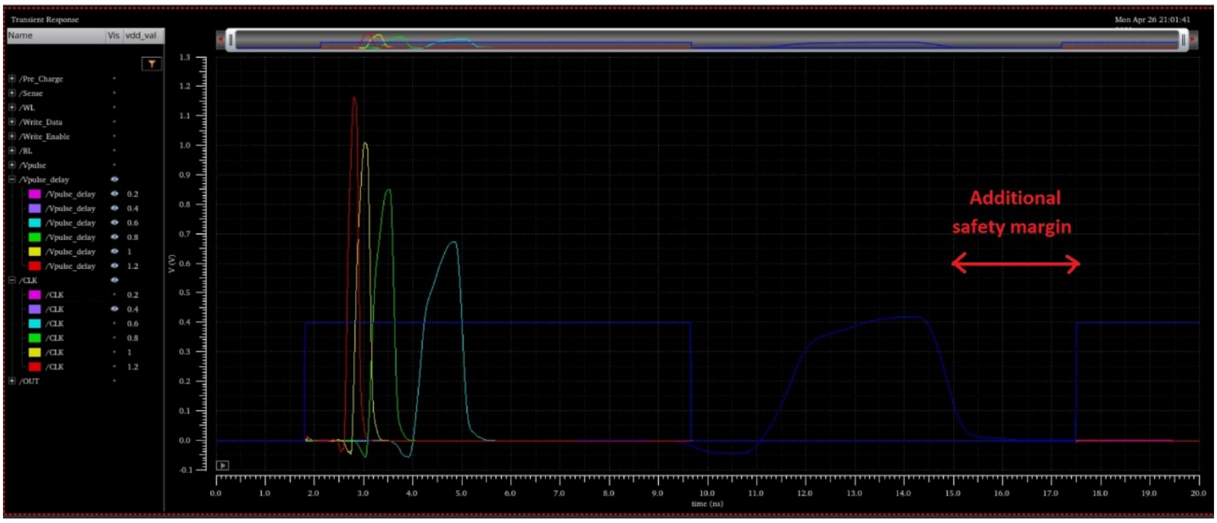


Figure 5.11: Parametric analysis of the delay element in SRAM memory for $V_{DD} \in [0,4V;1,2V]$.

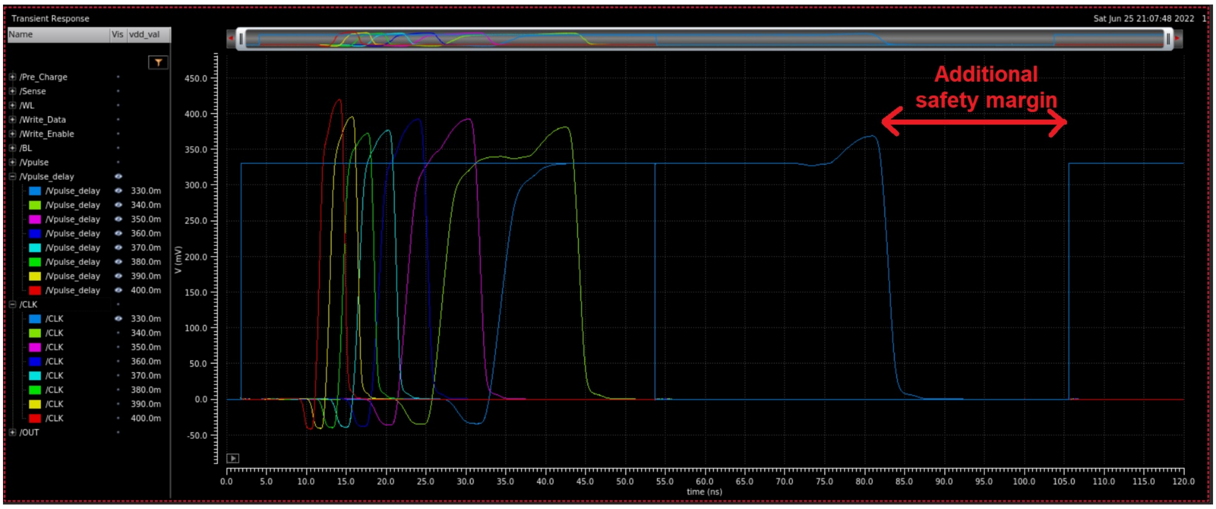


Figure 5.12: Parametric analysis of the delay element in SRAM memory for $V_{DD} \in [0,33V;0,4V]$.

Through figure 5.11 it is possible to see that the pulses generated in the sensor have been shifted to the right, as you would expect. The blue line represents the clock for a V_{DD} of 0.4V, and note that in this example an additional safety margin is used, which means that the sensor is not working on the imminence of signaling an error. In Figure 5.12 it is possible to observe in more detail, that for a minimum V_{DD} of 0.33V, the delay element is still capable of producing pulses that are quite wide and require a clock with a longer period, to maintain the same percentage of additional safety margin.

5.1.3 Flip-Flop

The Ultra-Low Power Performance Sensor features a last block referring to a D flip-flop, which on its output (OUT), signals ”1” for the detection of an error and ”0” for safety operation of the memory cell. This flip-flop detects an error when the signal generated at the output of the delay element block reaches the second rising edge of the clock. The architecture of the D flip-flop is shown in Figure 5.13.

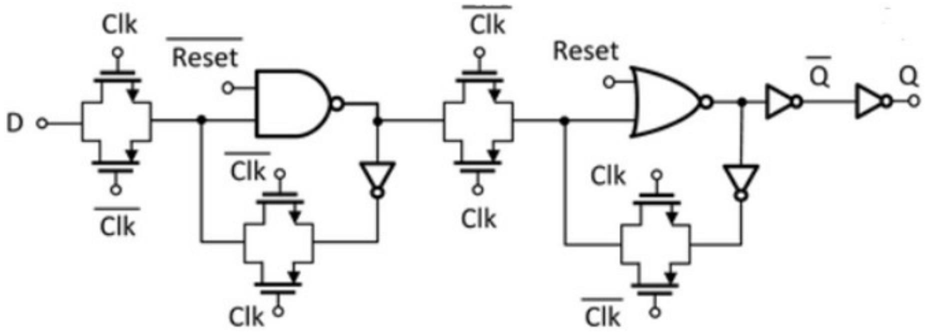


Figure 5.13: D flip-flop for Ultra-Low Power Performance Sensor.

Figure 5.13 represents a typical flip-flop architecture consisting of 4 transmission gates, 4 inverters, one NAND gate and one NOR gate that allow you to activate Reset when an error is detected to re-initialize signals. This flip-flop has the following inputs: the data (D), the clock (CLK) and the Reset. The output Q corresponds to the output of the sensor (OUT).

At the flip-flop input D it is necessary to connect an OR gate with inputs connected to the OUT and Vpulse_delay (figure 5.14), in order to, when an error is detected in the sensor, the output (OUT) remains at the logical value ”1”, until the Reset is activated. This will retain an error signal in the sensor for more than one clock cycle (in this case, until the reset is activated), and allows that corrective measures can be taken to change performance and avoid errors.

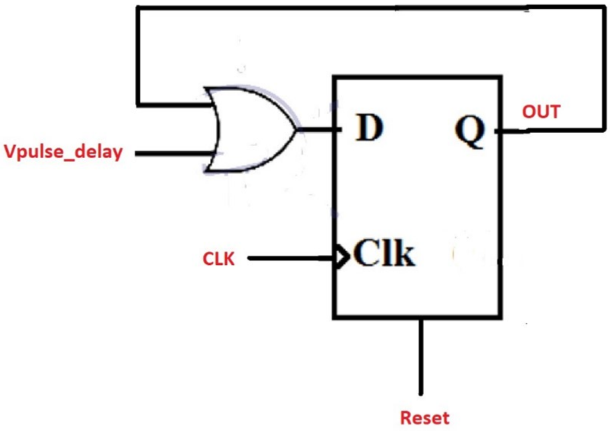


Figure 5.14: Flip-flop block.

In order to show the operation of the flip-flop block, two simulations were carried out for Vdd=0.8V. The first simulation is done with a clock period of 2.2ns (454MHz) and the second simulation with a clock period of 1.7ns (588MHz). Each simulation corresponds to a writing of the value ”0” on the SRAM memory cell.

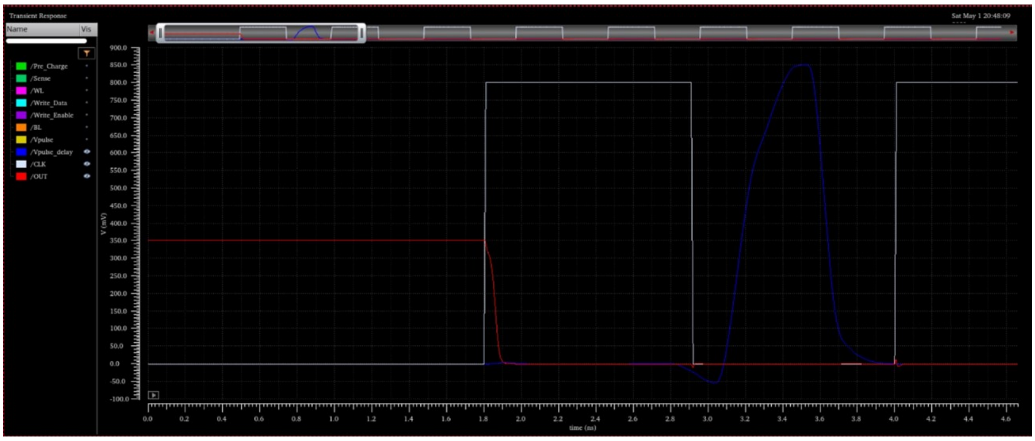


Figure 5.15: Flip-flop block operation in SRAM memory for V_{DD}=0.8V and clock_period=2.2ns.

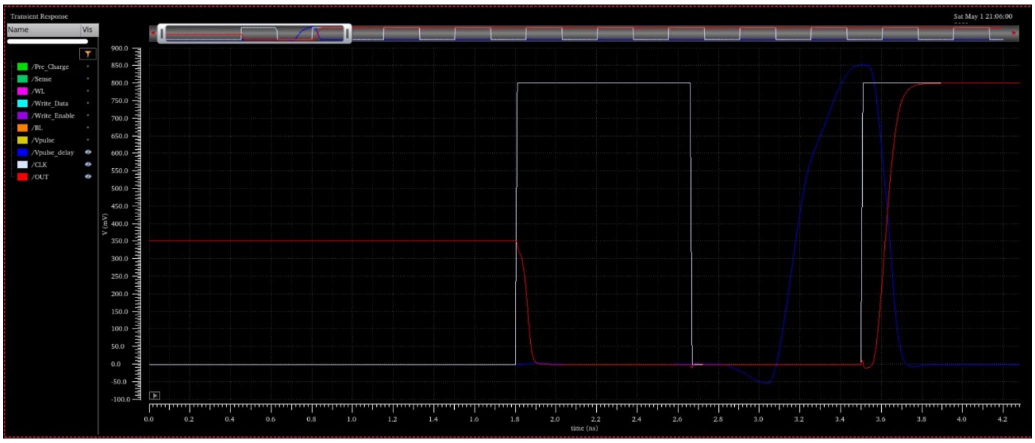


Figure 5.16: Flip-flop block operation in SRAM memory for V_{DD}=0.8V and clock_period=1.7ns.

Through figure 5.15, we can see that for a clock period (white line) large enough, the pulse generated on the sensor (blue line), does not reach the second rising edge of the clock and therefore the sensor output (red line) has the value of 0V, thus ensuring that this operation does not present problems. While in figure 5.16 the clock period is reduced and this time the pulse (blue line) reaches the second rising edge of the clock, implying that the sensor output (red line) will rise to 0.8V, meaning that for a lower clock period the operation is no longer safe and the sensor indicates an error. We can thus conclude that when there are PVTA variations, the width and pulse delay (Vpulse) will be increased, which leads to the pulse reaching the second rising edge of the clock, signaling the occurrence of error. On the other hand, when lowering the V_{DD}, the pulse also increases its delay and width, making it necessary to use a DVFS technique that allows you to change the clock frequency as the supply voltage decreases to ensure that there is always a safety margin in the circuit.

In figure 5.17 and 5.18 it is possible to observe that the flip-flop block continues to work for V_{DD} equal to 0.33V, detecting errors when the pulse reaches the second rising edge of the clock, i.e., the Ultra-Low Power Performance Sensor can operate for minimum supply voltages up to 0.33V.

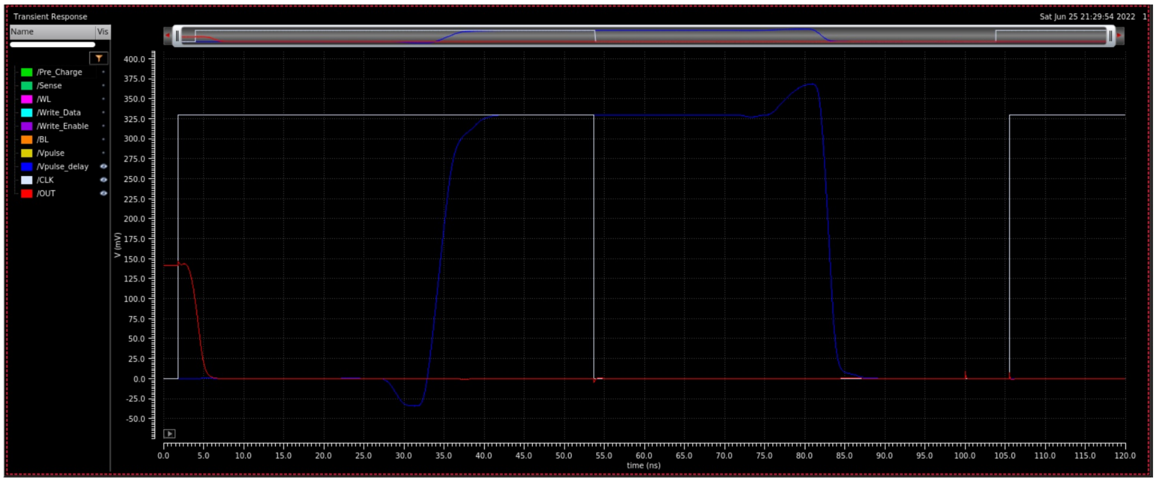


Figure 5.17: Flip-flop block operation for V_{DD}=0,33V and clock_period=103,75ns.

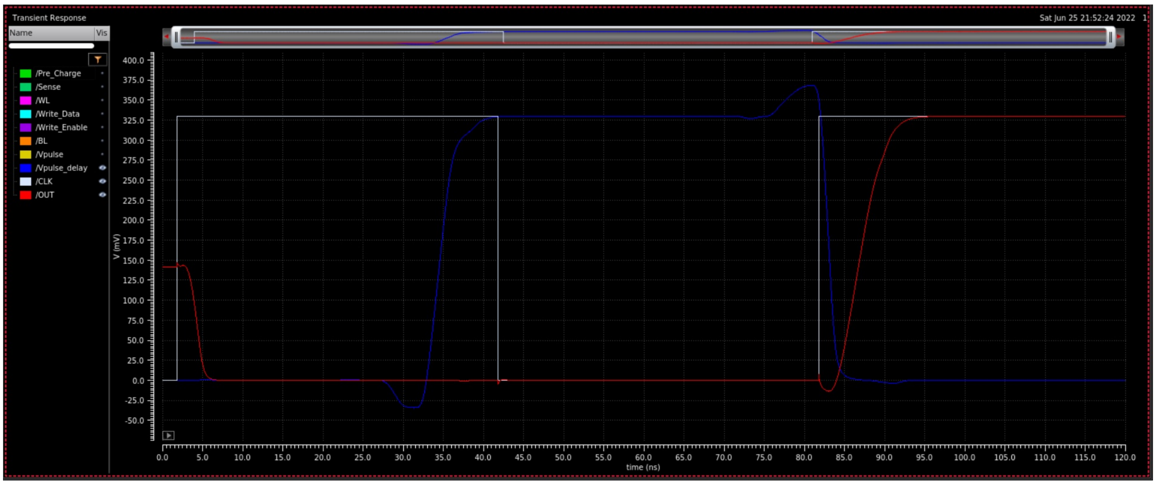


Figure 5.18: Flip-flop block operation for V_{DD}=0,33V and clock_period=80ns.

5.1.4 Complete Circuit

In this subsection is presented the complete circuit of the Ultra-Low Power Performance Sensor, using a version 1 transition detector (figure 5.19).

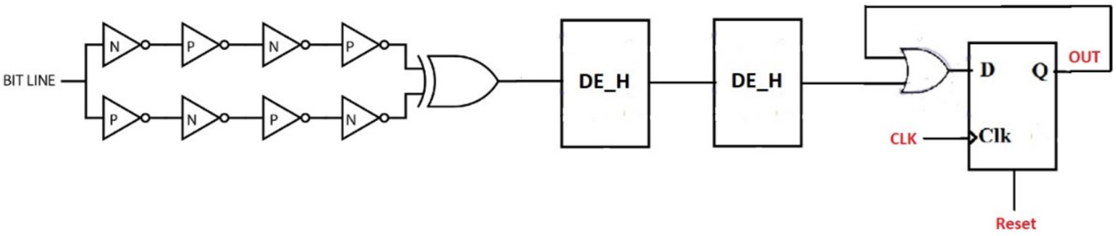


Figure 5.19: Complete circuit for Ultra-Low Power Performance Sensor – version 1.

As we can see the sensor consists of the transition detector block, two delay elements and a flip-flop block. This sensor signals signal degradation during Read/Write operations in a memory cell due to

	Delays						
Vdd	Signals to write in memory(Write)	Writing of signals (BL)	Use of written data	Sensor Delay (DE)	Clock period	Safety margin	Safety margin (%)
1,2 V	205,01 ps	235,76 ps	699,94 ps	1,09 ns	1,32 ns	0,39 ns	30,1 %
1 V	213,03 ps	243,01 ps	828,42 ps	1,34 ns	1,61 ns	0,51 ns	31,9 %
0,8 V	223,97 ps	253,79 ps	1,08 ns	1,83 ns	2,2 ns	0,75 ns	34,2 %
0,6 V	259,01 ps	289,49 ps	1,81 ns	3,23 ns	3,88 ns	1,42 ns	36,8 %
0,4 V	411,98 ps	690,69 ps	6,81 ns	13,07 ns	15,69 ns	6,26 ns	39,9 %
0,33 V	1405,02 ps	1914,92 ps	62,08 ns	81,27 ns	97,53 ns	19,2 ns	19,7 %

Table 5.1: Delays of the signals for write operation in SRAM cell.

PVTA variations. The sensor detects a transition in the bit line and generates a pulse proportional to the transition time. Due to the presence of delay elements, a delay is added to the pulse, to accommodate a similar delay to the circuit critical path. Then a flip-flop is used to detect the pulse when it reaches the second rising edge of the clock.

As already mentioned above, this sensor is compatible with the use of a DVFS technique to allow working reduced supply voltages. The purpose of this technique is to correspond for each V_{DD} value, a different frequency from the clock, so that the sensor works always at maximum performance (i.e., on the imminence of an error), ensuring that a safety margin remains for all V_{DD} levels.

To exemplify the operation of the DVFS technique, 6 simulations were performed for various V_{DD} values, during a writing operation of the value ”0” in a SRAM. Transistors with size WNmin=160nm and WPmin=600nm were used. Table 5.1 shows the delay values of the various signals for each V_{DD} and the clock period calculation, considering that there is an additional safety margin of 20% for a normal operation.

The various delays of the table are measured from the first rising edge of the clock to the variation of the respective signal, and the calculation of the clock period is equal to the delay of the Delay Sensor (DE), plus 20% of an additional safety margin (to consider that in a correct operation, the sensor does not produce a predictive error). In the table, it is also possible to observe that the safety margin increases when the V_{DD} is lowered, which means that the sensor becomes more cautious when working in worse conditions. Note that for V_{DD} =0.33V, the safety margin decreases due to the fact that the pulse delay is too high, which means that the normal operation of this sensor at the simulated clock frequency is for supply voltages of 0.4V (or it may work at 0.33V of V_{DD}, but a lower frequency should be used, to maintain a similar additional safety margin of 20%).

Next, two graphics of the simulations performed for V_{DD}=1.2V and V_{DD}=0.4V are shown.

Figures 5.20 and 5.21 show the sensor running at maximum performance for a normal operation, to maintain the additional safety margin of 20% of the clock period. The red line corresponds to the signal to write in memory signal, the yellow line to the transition of the bit line, the green line simulates the delay caused by the use of the written data (critical path), and the light blue line is the pulse signal at the output of the Delay Element block. The white lines represent the respective delays of each of the signals relative to the first rising edge of the clock. It is possible to observe that the delays of the various signals increase in size when the V_{DD} is smaller and so the clock period also increases.

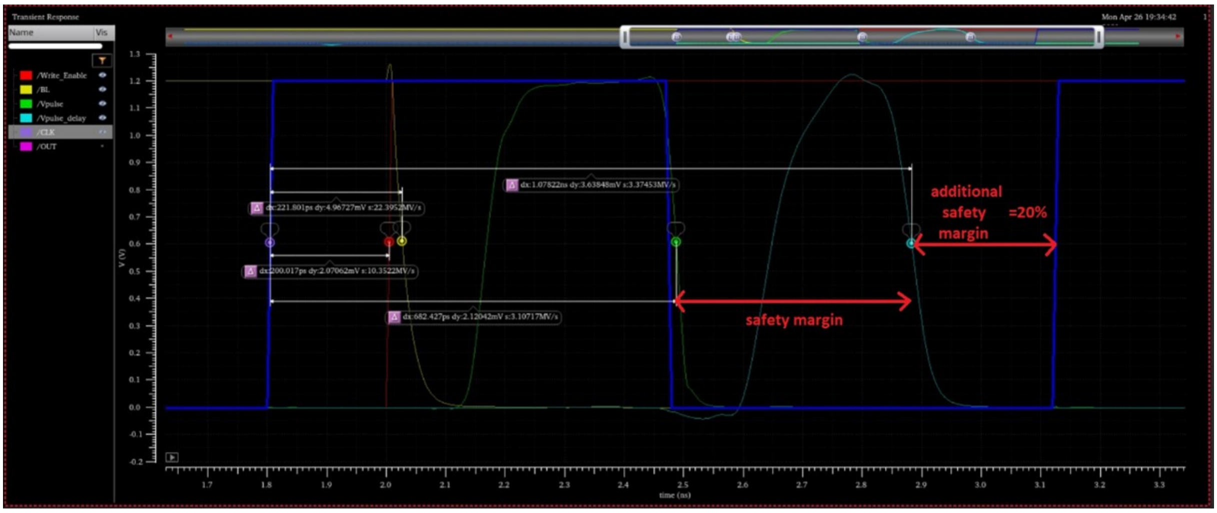

Transient Response
Mon Apr 26 19:34:42
Name
Vis
/Write_Enable
/BL
/Vpulse
/Vpulse_delay
/CLK
/OUT
dx:1.07822ns dy:3.63848mV s:3.37453MV/s
dx:221.801ps dy:4.96727mV s:22.3952MV/s
dx:200.017ps dy:2.07062mV s:10.3522MV/s
dx:682.427ps dy:2.12042mV s:3.10717MV/s
additional safety margin
=20%
safety margin
V (V)
time (ns)

Figure 5.20: Graphic of the signals delays for V_{DD}=1.2V.

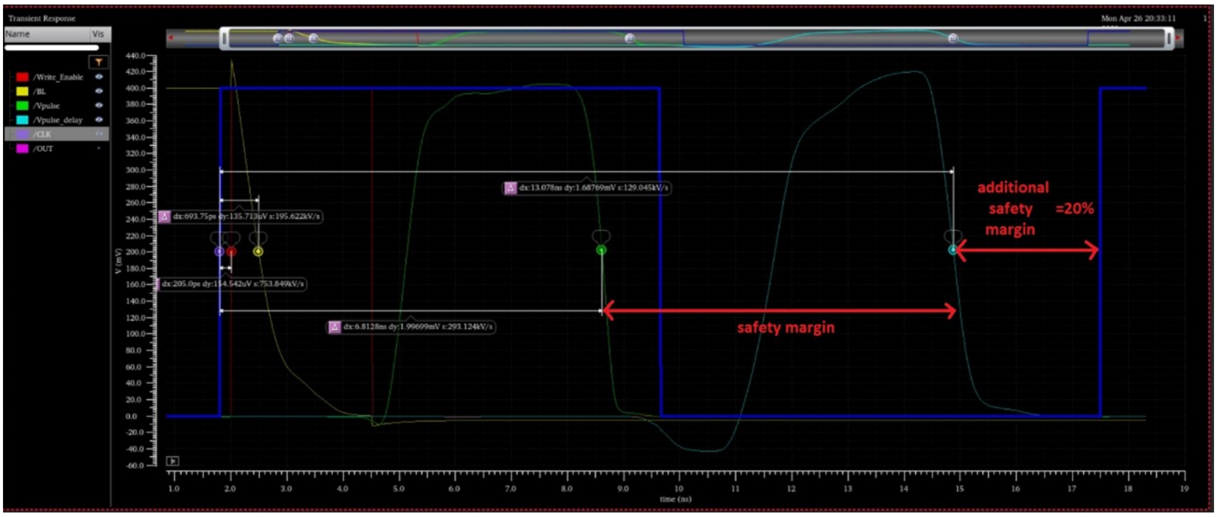

Transient Response
Mon Apr 26 20:33:11
Name
Vis
/Write_Enable
/BL
/Vpulse
/Vpulse_delay
/CLK
/OUT
dx:13.078ns dy:1.68769mV s:129.045kV/s
dx:693.75ps dy:135.713uV s:195.622kV/s
dx:205.0ps dy:154.542uV s:753.849kV/s
dx:6.8128ns dy:1.99699mV s:293.124kV/s
additional safety margin
=20%
safety margin
V (mV)
time (ns)

Figure 5.21: Graphic of the signals delays for V_{DD}=0.4V.

Regarding the complete circuit for sensor versions 2 and 3, Figure 5.22 presents the version 2, to monitor both bit lines in an SRAM memory cell initialized at V_{DD}, and Figure 5.23 presents the version 3, to monitor one bit line in a DRAM memory cell, initialized at V_{DD}/2.

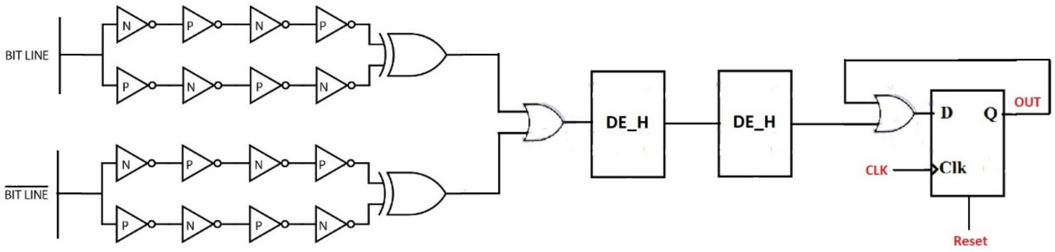


Figure 5.22: Complete circuit for Ultra-Low Power Performance Sensor – version 2.

In previous section 5.1.1, the simulations for the different versions of the transition detector were presented and we could see that similar pulses were generated. Hence, when considering the complete

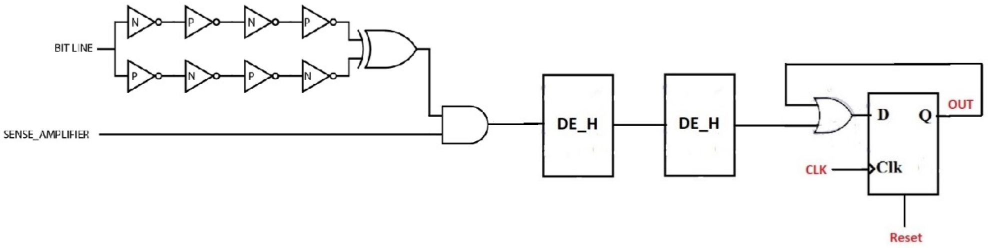


Figure 5.23: Complete circuit for Ultra-Low Power Performance Sensor – version 3.

circuit, apart from the pulse detection, everything else has the same implementation in all 3 versions. Therefore, the simulation results for complete sensor circuits for versions 2 and 3 are similar to version 1, already presented in this section and, to avoid repetitions and to simplify this document, these simulations will not be presented here, as the results would be the same.

5.2 Discussion and Analysis on Sensor Usability

The sensor presented in this chapter can be used in SRAM and DRAM memories and for any type of architecture of the memory cell, because the sensor works as long as transitions occur in the bit line. This sensor works for low supply voltages down to 0.33V, from which the memory cell stops working properly and the use of the sensor becomes inappropriate.

This sensor is quite versatile because it allows the designer to use different types of transition detector blocks, delay element blocks, and can be applied to different memory types. If higher sensitivity to bit line transitions is required, the designer can increase the number of inverters used in the transition detector to obtain wider pulses. If a higher safety margin is required, or if a higher critical path exists in the circuit, the designer can use more delay elements to increase the delay of the delayed pulse.

5.2.1 Sensor placement in the memory

In section 2.1, the SRAM and DRAM memories' architectures were presented and, as we can see in the description, the sense amplifiers are placed in the columns of a memory, connected to the bit lines. This performance sensor for memories is also connected to the bit lines, and each sensor monitors all the cells connected in a column. In fact, by selecting the word line, only one memory cell in each column will be active at a time, allowing to monitor each cell individually. Of course, that happens if all the columns (bit lines) have a sensor connected to it, but this may not be the solution for all cases.

If we consider a big memory, for example, 4 GB (which is a normal size in DRAMs), a typical number for the memory arrangement is ([30]):

- Number of Row Address bits: A0-A15 = 16 bits
  - Total number of row = 2^{16} = 64K
- Number of Column Address bits: A0-A9 = 10 bits

  - Number of columns per row = 1K

- Width of each column = 4 bits
- Number of Bank Groups = 4
- Number of Banks = 4
- Total DRAM Capacity = Num.Rows x Num.Columns x Width.of.Column x Num.BankGroups x Num.Banks = 64K x 1K x 4 x 4 x 4 = 4Gb

This means that, for each bank, 1024 sensors are needed, with 65536 cells connected to each bit line. In other words, and considering that all the bit lines are monitored, each sensor would monitor 65536 cells. For the entire 4GB memory, 65536 sensors are needed, which is less than 0.002% of sensors per memory cell.

Although as the memory size grows and more cells may be monitored by each sensor, the number of sensors used may be reduced even more if we consider that, statistically, we may obtain similar results of the monitoring procedure with a random representation of the memory and using fewer sensors and fewer monitored cells. Nevertheless, this approach needs to be validated, and its work is not in the scope of this thesis, and this should be pursued in future work, not only to validate this approach but also to identify the minimum ratio of monitored cells per total cells that can correctly represent the memory usage.

Anyway, the new performance sensor presented in this work may be used to monitor all the cells in the memory, or to monitor a representative set of cells in the memory.

5.2.2 Sensor usage in an ultra-low-power DVFS strategy

Typically, a DVFS strategy is used when power reduction is needed. Considering the case of ultra-low-power strategies, this means that subthreshold voltage levels can be used in the power-supply. Hence, an ultra-low-power DVFS strategy refers to a technique that reduces the power-supply voltage V_{DD} to subthreshold levels to drastically reduce power consumption, and, consequently, the clock frequency f_{CLK} should also be reduced, not only to increase power savings, but also to avoid errors due to V_{DD} reduction. Therefore, pairs of V_{DD} and f_{CLK} values are normally stored in a table and used in a circuit operation, and they are chosen carefully to avoid performance errors. This approach is the typical DVFS approach, because normally performance in a circuit is affected by many parameters and it's not easy to monitor each parameter individually and to know exactly how the circuit is affected by the change of one (or many).

However, with this new performance sensor for CMOS memory cells, the sensor can monitor the performance of the cell, regardless of the parameter (or parameters) that are affecting its behavior. We identified PVTA variations as the most important ones that change memories' performance, but in fact it may monitor any other parameter that changes the cell timing response. Since the sensor monitors the reading/writing operations in the bit lines, it may monitor errors in the cell, but also in the memory circuitry (like the sense amplifier) that can change the reading/writing operations' timings.

Moreover, as the sensor uses a minimum safety margin defined by design, it can detect the unsafe transitions locally, where the errors occur in the memory and where it can affect other circuits connected to the memory. This way, the sensor can be used to control dynamically an ultra-low-power DVFS strategy, because it can search for the optimum V_{DD} / f_{CLK} pair that places memory operation working at the minimum safety margin, and thus optimizing circuit operation.

Also, the optimization strategy can work both ways, that is, both variables can be controlled and optimized: V_{DD} and clock frequency. It can be established a fixed V_{DD} and f_{CLK} is changed to search for the minimum safety margin; or it can be established a fixed clock frequency and the V_{DD} is changed to search for the minimum safety margin; or even a combine strategy is used to optimize circuit operation.

Nevertheless, the new sensor can be a key factor to dynamically control memory operation and allow working with a minimum safety margin, but still avoiding errors.

5.2.3 Local Sensor vs Global Sensor

In other performance sensors described in literature ([19]), there are typically two approaches used to monitor circuit operation: use local sensors, to monitor errors locally, where they happen; or use global sensors, by monitoring typically a dummy copy of the circuit and extrapolate to the real circuit the monitoring results of the copy.

When using local sensors, sensors are placed locally where the errors occur, which in this case is in the bit lines of the memory, where the reading and writing of the memory occurs. But in our case, and like other performance sensor approaches ([19]), the sensor needs to be activated with a reading and/or writing procedure that triggers a change in the bit line. This means that, even though the sensor is always active, if the reading and writing procedure does not occur, the sensor monitoring procedure is not activated and the memory performance is not tested. This could be a problem if we consider, for example, that a cell can have transistors in stress mode condition and aging, not being tested during all the aging degradation, and when it needs to be used on a reading procedure, it fails, and a real error can happen. So, the monitoring procedure is dependent of the cell usage by the user, which can be a problem, as explained. We can argue that this degradation may be observed in a different cell that may be activated, but in reality, this cannot be guaranteed. Moreover, they are more intrusive to the circuit and, therefore, they are more difficult to install in a circuit (in this case in the memory), because the memory needs to be design along with the sensors. In conclusion, local sensors have the advantage of being able to monitor the real bit lines, cells and operations in the memory, but they have the disadvantage of not having the guarantee of being activated, depending on the cell usage by the user, and are more difficult to use.

When using Global sensors, the circuit design (in our case, the memory) is made separate from the sensor design, and one copy of the sensor is used with a copy of the circuit, which in our case would be a copy of one, or more, memory cells. Because the real memory is not connected to the sensor, but a sample copy, this means that sensor activation can be forced at any time and does not depend on the real memory usage. Normally, the sensor monitors periodically the copied memory cell, which mimics the real memory behavior, and extrapolates the result to the entire memory. However, the copied cell may age differently than the real memory, so the extrapolation of the sensor output in the dummy cell to the real memory may be a wrong assumption. Even in the presence of local PVTA degradations in a circuit, for example, process variations are different in each transistor, temperature hotspots, V_{DD} fluctuation, or aging differently in different memory cells, although we can use additional safety margins in the global sensor to account with unpredictable variations when compared with the real memory, we do not have any guarantee that the dummy cell will represent correctly the real memory. Nevertheless, global sensors are less intrusive and are normally well adopted by industry, due to the fact that are easy to implement (they do not interfere directly with normal circuit design). In conclusion, global sensors have the advantages of being easy to activate when needed and are easy to use in a circuit, but they have the disadvantage of not having the guarantee of a correct circuit monitoring, because they monitor a dummy copy of the circuit.

Interestingly, as done in other previous works for logic circuits ([19]), both strategies can be used together to increase reliability in the memory and make use of the advantages of both strategies. In our new performance sensor for CMOS memory cells, local sensors can be used to detect PVTA degradations when the bit lines change, and the sensors are activated. This information can give a real information about the memory. But at the same time, a global sensor may be install and guarantee a periodic activation to overcome the fact that local sensors may not be periodically activated. Moreover, off-line tests on local sensors can be used to adjust and tune global sensors, as done in ([19]). Note, that the local sensor can be used not exhaustively on all memory cells, but only on a few bit lines, using a statistical distribution.

In this thesis, a new local sensor was presented, and it can be used, as explained, to defined a global sensor based on the local sensor architecture. However, this global sensor design and the sensor placement in a memory circuit is out of the scope of this thesis and will be addressed for future work.

5.2.4 Resume of Sensor Characteristics

The Ultra-Low Power Performance Sensor for memories has several interesting characteristics, as stated bellow, some are advantages, other disadvantages of the new sensor:

- Simple and robust design, with increased reliability when conditions are worse.

Sensor design and operation is based on delays and on increase delays when PVTA variations get worse. The delays in the sensor operation depend mostly on the delay element, which is a simple buffer with small transistors, and the smaller the transistors are, more delay is added to the signal and more sensitive the circuit is to PVTA variations. This means a simpler design produces more sensitive sensors and more cautious operation on predicting errors in the memory operation. In other words, if operating conditions are worse (more aging, lower V_{DD}, higher temperature), the sensor becomes even more cautious, preventing errors more easily.

- The memory sensor is independent of the architecture of the memory cells.

This sensor is independent on the architecture of the memory cell. As the sensor works by monitoring the bit lines, the memory cell architecture does not influence on sensor usage. For example, the typical SRAM cell is the 6T (six transistors) cell, but for low-power memory cells, versions of 8T, 10T, 11T, etc. can be used. On the other hand, any aging that exists in the memory cell or even in the sense amplifier, ends up being reflected in the speed of transition of the bit line, so this sensor detects errors coming from the various structures of the memory. Moreover, it also has the advantage of being able to be used, with slight changes, both in SRAM and DRAM memories.

- Monitors performance degradation in memory usage, regardless the parameter that produces the change in performance.

The sensor presented in this chapter is a performance sensor that detects errors regardless of their origin. Process, Supply Voltage, Temperature and Aging, are the most common parameters, but any other that may affect the performance of the memory may be monitored with this performance sensor. This is because the error is reflected only in the speed of transition of the bit line.

- Multipurpose sensor

Regarding sensor usability and placement, this approach is very versatile, because it can be used as local sensor, monitoring all the bit lines or monitoring a random sample of cells, but it also can be used to define a new global sensor, or combine both approaches, as done in other design strategies.

- Sensibility is defined by design and cannot be changed during lifetime.

The Ultra-Low Power Performance Sensor has the disadvantage that its sensibility is based on the delays defined by design, which cannot be changed during circuit lifetime. This means that additional safety margins have to be used to account with unpredictable changes in the manufacturing process. Moreover, the real memory chips should be individually tested to understand how the sensor works for the different V_{DD}, and what is the delay map of the circuit. Additionally, if a DVFS strategy is used, the pairs of V_{DD} and clock frequency will have to be fine-tuned almost chip by chip. These problems raise the need for a sensibility tuning for online operation of the sensor, and this should be addressed in future work.

- The use of large clock periods imposes the use of many delay elements

Another disadvantage of this sensor is that if we use a clock with a large period, it may be necessary to use many delay elements to be able to increase the delay and produce the pulse near the second rising edge of the clock. This fact considerably increases sensor area and memory circuit overhead.

- Large area overhead for small size memories

The existence of a flip-flop (which is a 18 transistors' cell), along with the possibility of use many delay elements and unbalanced inverters (in the transition detector block), makes the sensor a relatively large block. For a bigger size memory, the sensor overhead may be negligible, but if a small memory is used, sensor overhead may be prohibitive.

Chapter 6

Implementation Layouts

In this section, the layouts for the memories and performance sensors are presented. First, we present the layout implementation of a sample memory, in this case an SRAM memory. Then, the layout for the Ultra-Low Power Performance Sensor is also presented, highlighting each constituent block of the sensor. Finally, the layout for the complete circuit of a 1-bit SRAM memory with the sensor.

The analysis of the layouts allows us to understand the dimensions and surface area of the circuits and their blocks. To perform these layouts, Cadence software and UMC130nm technology were used.

6.1 1-Bit SRAM

This section shows the layout of a 1-bit SRAM memory consisting of four blocks (SRAM cell, sensor amplifier, the pre-charge and equalizer circuitry, the write circuitry). Finally, a layout is also shown with the four blocks together.

Transistors with sizes were used for this technology: $L_{NMOS} = 0,4\mu m$, $W_{NMOS} = 0,5\mu m$, $L_{PMOS} = 0,4\mu m$ and $W_{PMOS} = 1,5\mu m$.

6.1.1 SRAM Cell

Figure 6.1 shows the layout of the SRAM cell. Through the figure it is possible to observe in the middle the cross-coupled CMOS inverters and the two NMOS access transistors. The top track is the V_{DD} and the bottom track is the ground (gnd). The light blue connections are in metal 1, in yellow they are in metal 2 and white in metal 5. The bit line (BL) is at the top and the complementary bit line (BLn) is located below. This circuit has as input the word line (WL).

The SRAM cell characteristics are:

– Width: 7 μm (700 lambda);

– Height: 11,8 μm (1180 lambda);

– Surface area: 82,6 μm^2.

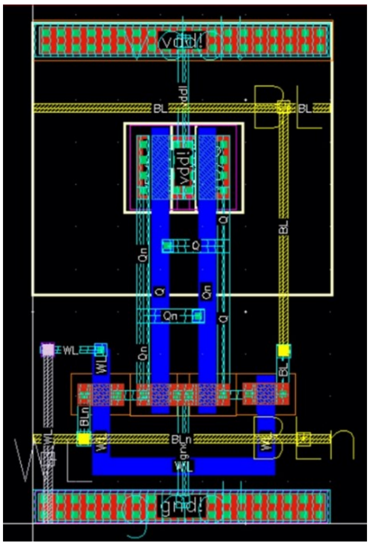


Figure 6.1: SRAM cell layout.

6.1.2 Sense Amplifier

The sense amplifier, provides amplification of small signal differences between the bit lines, responding with a full swing signal to guarantee the usage of the correct logic levels. Through figure 6.2 it is possible to see the latch formed by cross-coupling two CMOS inverters that are connected to the bit line (BL) and the complementary bit line (BLn). It is also possible to observe that there is an NMOS transistor (below) connected to the Sense signal and a PMOS transistor (above), connected to the complementary Sense signal. The complementary Sense signal also requires an inverter that appears on the right side in the layout. The top track is the V_{DD} and the bottom track is the ground (gnd). The light blue connections are in metal 1, in yellow they are in metal 2 and white in metal 5. The bit line (BL) is at the top and the complementary bit line (BLn) is located below. This circuit has as input the Sense signal.

The sense amplifier characteristics are:

– Width: 11 μm (1103 lambda);

– Height: 11,8 μm (1180 lambda);

– Surface area: 129,8 μm^2.

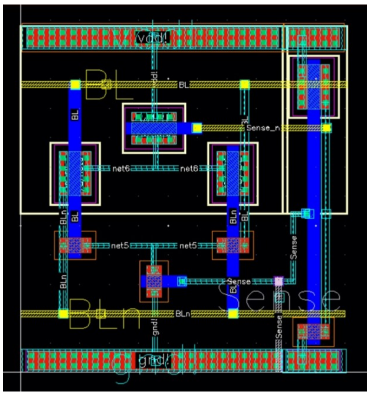


Figure 6.2: Sense amplifier layout.

6.1.3 Pre-Charge and Equalizer

Through figure 6.3, we can observe that the layout of the pre-charge and equalizer circuitry consists of two PMOS transistors, one of them connected to the bit line (BL) and the other connected to the complementary bit line (BLn) and with both gates connected to the Pre_Charge input signal. This circuit pre-charges and equalizes both bit lines to V_{DD}. The top track is the V_{DD} and the bottom track is the ground (gnd). The light blue connections are in metal 1, in yellow they are in metal 2 and white in metal 5. The bit line (BL) is at the top and the complementary bit line (BLn) is located below. This circuit has as input the Pre_Charge signal.

The pre-charge and equalizer characteristics are:

– Width: 5 μm (500 lambda);

– Height: 11,8 μm (1180 lambda);

– Surface area: 59 μm^2.

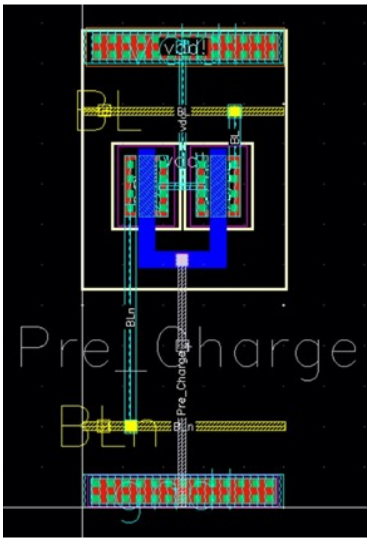


Figure 6.3: Pre-charge and equalizer layout.

6.1.4 Write Circuitry

The write circuitry, forces the bit line with the logic values intended to write on the SRAM cell (logic “0” or “1”). Figure 6.4 shows that the write circuitry layout consists of two NMOS transistors with the gates connected to the input signal Write_Enable and the source connected to the bit line (BL) and the complementary bit line (BLn). Figure 6.4 also shows two inverters that are connected to the input signal Write_Data. On the Write_Data signal are placed the bits to send to the SRAM cell, while the Write_Enable activates the NMOS transistors that send the values to the bit lines. In the layout the top track is the V_{DD} and the bottom track is the ground (gnd). The light blue connections are in metal 1, in yellow they are in metal 2 and white in metal 5. The bit line (BL) is at the top and the complementary line bit (BLn) is located below. This circuit has as input the Write_Data signal and the Write_Enable signal.

The write circuitry characteristics are:

– Width: 10 μm (1000 lambda);

– Height: 11,8 μm (1180 lambda);

– Surface area: 118 μm^2.

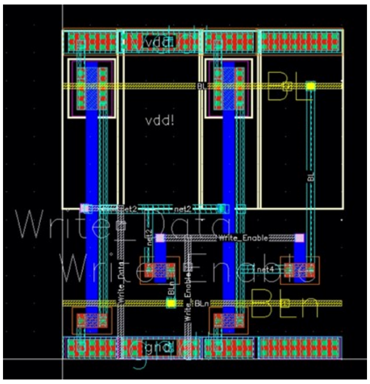


Figure 6.4: Write circuitry layout.

6.1.5 Complete 1-Bit SRAM

Figure 6.5 shows the complete 1-bit SRAM layout, consisting of the four blocks already mentioned. In the layout the top track is the V_{DD} and the bottom track is the ground (gnd). The light blue connections are in metal 1, in yellow they are in metal 2 and white in metal 5. The bit line (BL) is at the top and the complementary bit line (BLn) is located below. This circuit has five inputs (WL, Sense, Pre_Charge, Write_Data and Write_Enable).

The complete 1-bit SRAM characteristics are:

– Width: 32,7 μm (3270 lambda);

– Height: 11,8 μm (1180 lambda);

– Surface area: 385,9 μm^2.

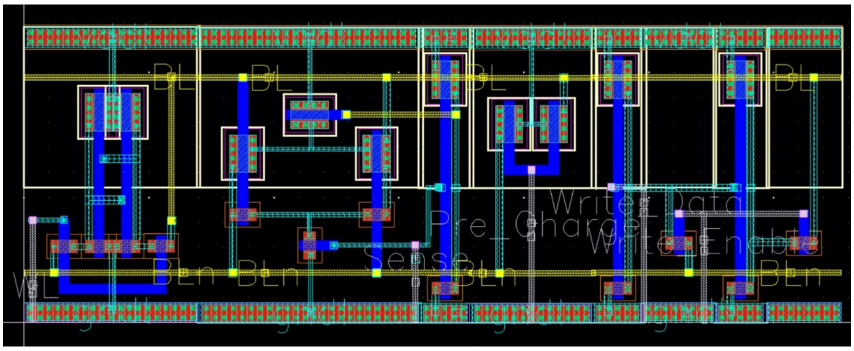


Figure 6.5: Complete 1-bit SRAM layout.

6.2 Ultra-Low Power Performance Sensor

In this section the layout for each block that constitutes the Ultra-Low Power Performance Sensor (transition detector, delay element, flip-flop) is presented. Finally, a layout is also shown for the three blocks together (the complete sensor).

6.2.1 Transition Detector

In Figure 6.6 it is possible to observe the layout of the transition detector block containing on the left side two paths with four inverters each, alternating a more conductive PMOS and NMOS MOSFET, in a total of 8 inverters. Each inverter has transistors with minimum sizes (for the NMOS transistor is WNmin=160nm and for the PMOS is WPmin=600nm). The design of the more conductive transistors was done using a 3-finger gate to occupy the space more efficiently on the cell.

On the right side of the layout is the XOR gate implemented with pass-transistor logic, being responsible to generate a pulse as a transition occurs in the bit line. The size of XOR gate transistors' for this technology are: $L_{NMOS} = 0,4\mu m$, $W_{NMOS} = 0,5\mu m$, $L_{PMOS} = 0,4\mu m$ and $W_{PMOS} = 1,5\mu m$. In the layout, the top track is the V_{DD} and the bottom track is the ground (gnd). The light blue connections are in metal 1 and yellow are in metal 2. This circuit has as input the bit line (BL) and as output the Vsense signal.

The characteristics of the transition detector are:

– Width: 32 μm (3196 lambda);

– Height: 11,8 μm (1180 lambda);

– Surface area: 377,6 μm^2.

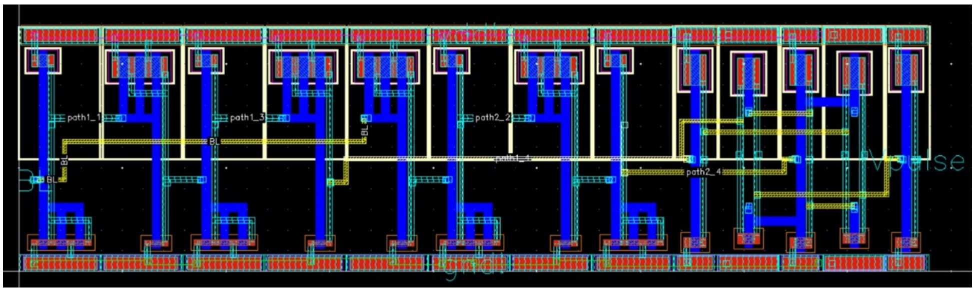


Figure 6.6: Transition detector layout.

6.2.2 Delay Element

The delay element block provides an additional time delay, to generated pulses from the transition detector. Figure 6.7 shows that this block contains two delay elements of type H (DE_H). Each delay element, consists of two inverters, connected to two transmission gates. Transistors with the following sizes were used for this block: $L_{NMOS} = 0,4\mu m$, $W_{NMOS} = 0,5\mu m$, $L_{PMOS} = 0,4\mu m$ and $W_{PMOS} = 1,5\mu m$. In the layout the top track is the V_{DD} and the bottom track is the ground (gnd). The connections in light blue are in metal 1. This circuit has as input the Vpulse signal and as output the signal Vpulse_delay.

The delay element characteristics are:

– Width: 15,3 μm (1532 lambda);

– Height: 11,8 μm (1180 lambda);

– Surface area: 180,5 μm^2.

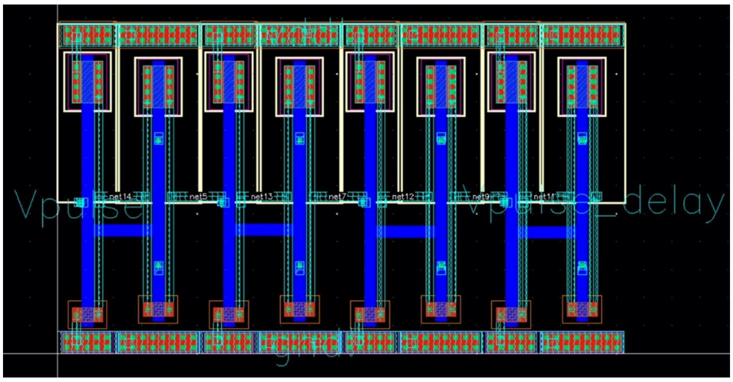


Figure 6.7: Delay element layout.

6.2.3 Flip-Flop

In Figure 6.8 it is possible to observe the layout of the flip-flop block containing from the left side to the right side, an OR gate, two transmission gates, one NAND gate, one inverter, again two transmission gates, one NOR gate and one inverter. Transistors with the following sizes were used for this technology: $L_{NMOS} = 0,4\mu m$, $W_{NMOS} = 0,5\mu m$, $L_{PMOS} = 0,4\mu m$ and $W_{PMOS} = 1,5\mu m$. In the layout the top track is the V_{DD} and the bottom track is the ground (gnd). The light blue connections are in metal 1, in yellow they are in metal 2, the green is in metal 3, purple in metal 4 and white in metal 5. This circuit has as inputs: the Vpulse_delay signal, the Clock, the Reset; and as output: the OUT signal.

The characteristics of the flip-flop are:

– Width: 25 μm (2499 lambda);

– Height: 11,8 μm (1180 lambda);

– Surface area: 295 μm^2.

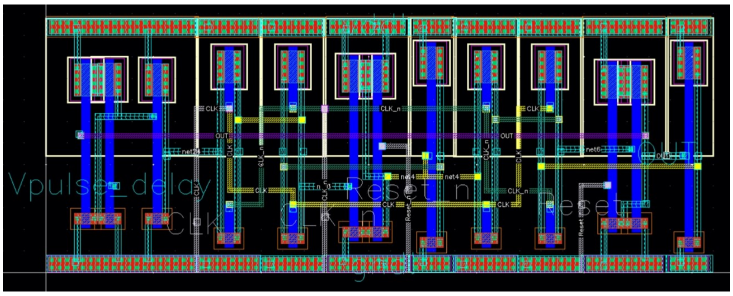


Figure 6.8: Flip-flop layout.

6.2.4 Complete Sensor

Figure 6.9 shows the complete layout of the Ultra-Low Power Performance Sensor with the three building blocks (transition detector, delay element and flip-flop). In the layout the top track is the V_{DD} and the bottom track is the ground (gnd). The light blue connections are in metal 1, in yellow they are in metal 2, the green is in metal 3, purple in metal 4 and white are in metal 5. This circuit has as inputs: bit line (BL), Clock, Reset; and as output: the OUT signal.

The characteristics of the complete sensor are:

– Width: 72 μm (7200 lambda);

– Height: 11,8 μm (1180 lambda);

– Surface area: 849,6 μm^2.

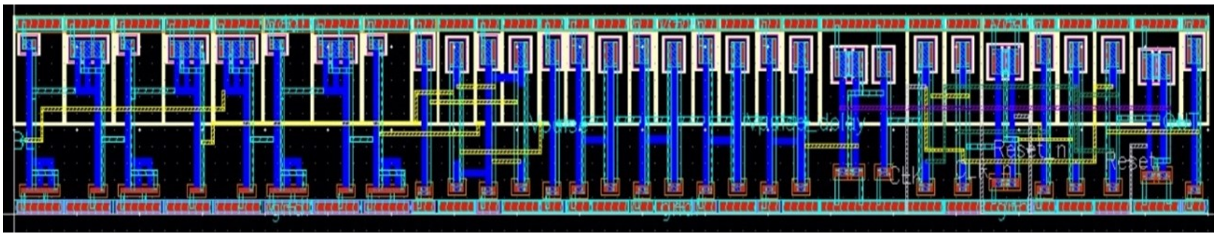


Figure 6.9: Complete sensor layout.

6.3 1-Bit SRAM With Ultra-Low Power Performance Sensor

To analyze the total size of the circuit, figure 6.10 shows the layout of the 1-bit SRAM memory (left side) along with the Ultra-Low Power Performance Sensor (right side). In the layout the top track is the V_{DD} and the bottom track is the ground (gnd). The light blue connections are in metal 1, in yellow they are in metal 2, the green is in metal 3, purple in metal 4 and white are in metal 5. This circuit has as inputs the signals: Word Line, Sense, Pre_Charge, Write_Data, Write_Enable, Clock, Reset; and as output: the OUT signal.

The characteristics of the 1-bit SRAM with Ultra-Low Power Performance Sensor are:

– Width: 104,6 μm (10455 lambda);

– Height: 11,8 μm (1180 lambda);

– Surface area: 1234,3 μm^2.

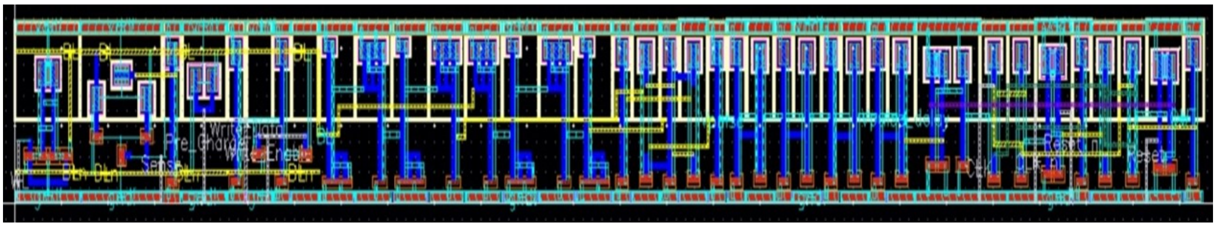


Figure 6.10: 1-bit SRAM plus Ultra-Low Power Performance Sensor layout.

6.4 64-Bit SRAM With Ultra-Low Power Performance Sensor

To get a more realistic idea of the sensor size, figure 6.11 shows the Ultra-Low Power Performance Sensor layout (right side) applied to a 64-bit SRAM memory (left side), with an 8 x 8 arrangement (8 word lines, 8 columns). In the layout the top track is the V_{DD} and the bottom track is the ground (gnd). The light blue connections are in metal 1, in yellow they are in metal 2, the green is in metal 3, purple in metal 4 and white are in metal 5. This circuit has as inputs the signals: Word Line, Sense, Pre_Charge, Write_Data, Write_Enable, Clock, Reset; and as output: the OUT signal.

The characteristics of the 64-bit SRAM with Ultra-Low Power Performance Sensor are:

– Width: 154 μm (15400 lambda);

– Height: 88,8 μm (8880 lambda);

– Surface area: 13675,2 μm^2.

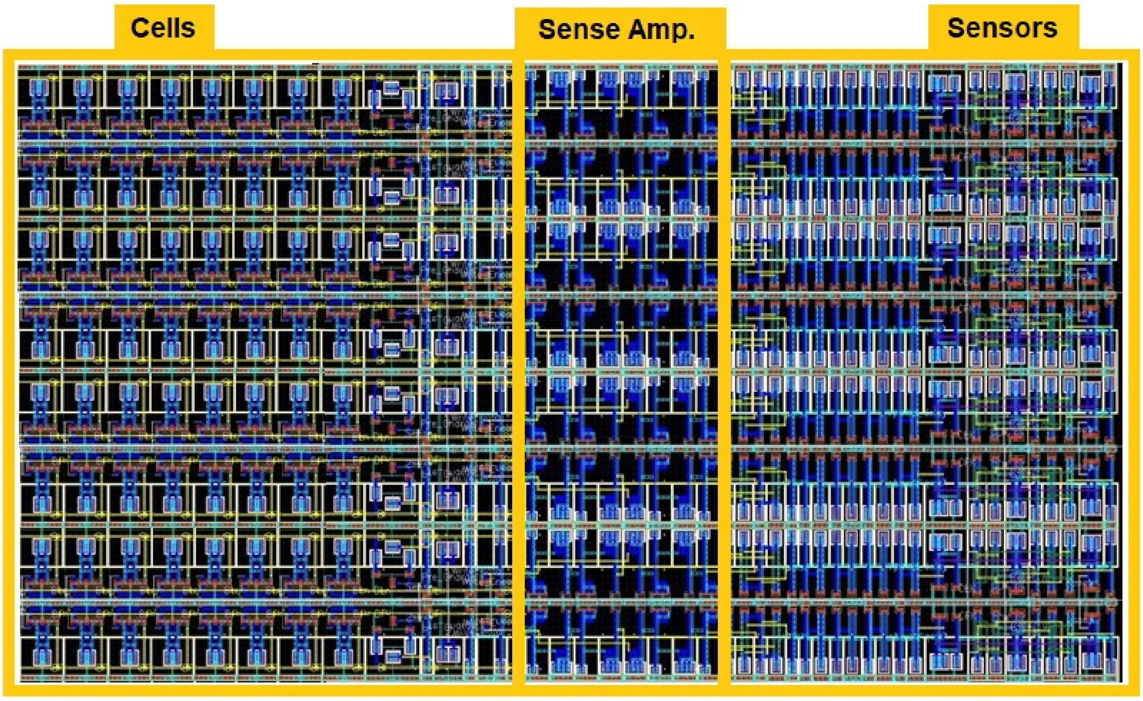


Figure 6.11: 64-bit SRAM plus Ultra-Low Power Performance Sensor layout.

6.5 Design Properties for Various SRAM Sizes

To analyze the size of the Ultra-Low Power Performance Sensor in relation to the total area (memory + sensor) table 6.1 was developed, based on the areas of the 1-bit and 64-bit memories presented, and calculating the required areas for bigger memories. This table shows the sensor areas in comparison with the memory areas, for different memory capacities. Note that data for SRAM memories greater than 64-bit were extrapolated but the same proportionality coefficient was maintained.

Through table 6.1 it is possible to observe that for low-capacity memories, the area overhead is significant, but for memories with higher capacity, especially with more cells per column (or more word lines in each column), the percentage of the sensor area decreases, as is the case of 64k-bit SRAM memory in which the sensor area represents only 3% of the total area. Note that we can still decrease the area overhead and complexity, because instead of using a sensor for each memory cell, we can use fewer sensors just monitoring a subset of the BL, assumed to represent the entire memory behavior.

SRAM Capacity	Arrangement lin x col	SRAM Area (μm^2)	Sensor Area (μm^2)	Total Area (μm^2)	Sensor Area (%)
1-bit	1 x 1	574,7 μm^2	659,6 μm^2	1234,3 μm^2	53,4 %
64-bit	8 x 8	9121,4 μm^2	4553,8 μm^2	13675,2 μm^2	33,3 %
1k-bit	32 x 32	94358,9 μm^2	20015,5 μm^2	114374,4 μm^2	17,5 %
64k-bit	256 x 256	5209505,3 μm^2	161118,7 μm^2	5370624 μm^2	3 %
16k-bit	2k x 8	1277040,4 μm^2	4698,8 μm^2	1281739,2 μm^2	0,39 %
512k-bit	16k x 32	40751808,9 μm^2	20178,3 μm^2	40771987,2 μm^2	0,05 %

Table 6.1: Area Design Properties for sensor plus memory.

6.6 DRAM cell

Figure 6.12 shows the layout of the DRAM cell. Through the figure it is possible to observe the storage capacitor that occupies most of the area at the top and the NMOS access transistor at the bottom. The top track is the V_{DD} and the bottom track is the ground (gnd). The light blue connections are in metal 1, in yellow they are in metal 2 and white in metal 5. The bit line (BL) is at the top. This circuit has as input the word line 1 (WL1).

The DRAM cell characteristics are:

– Width: 11,6 μm (1160 lambda);

– Height: 11,8 μm (1180 lambda);

– Surface area: 136,9 μm^2.

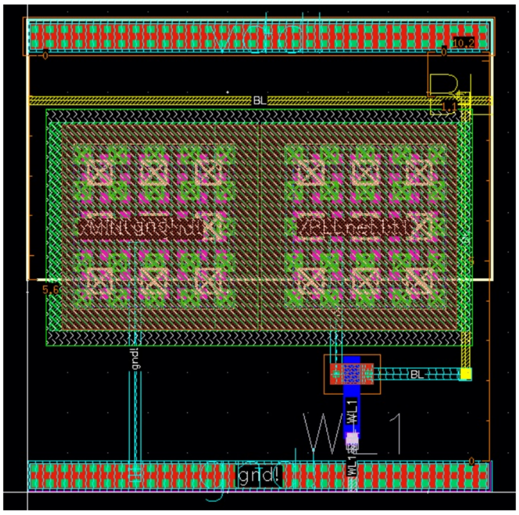


Figure 6.12: DRAM cell layout.

6.7 1-Bit DRAM With Ultra-Low Power Performance Sensor

This section shows the 1-bit DRAM (left side) together with the Ultra-Low Power Performance Sensor (right side) in order to compare with section 6.3 regarding a SRAM memory. From figure 6.13 it is possible to observe that the top track is the V_{DD} and the bottom track is the ground (gnd). The light blue connections are in metal 1, in yellow they are in metal 2, the green is in metal 3, purple in metal 4 and white are in metal 5. This circuit has as inputs the signals: Word Line 1 (WL1), Word Line 2 (WL2), Sense, Pre_Charge, Write_Data, Write_Enable, Clock, Reset; and as output: the OUT signal.

The characteristics of the 1-bit DRAM with Ultra-Low Power Performance Sensor are:

– Width: 143,6 μm (14362 lambda);

– Height: 11,8 μm (1180 lambda);

– Surface area: 1694,5 μm^2.

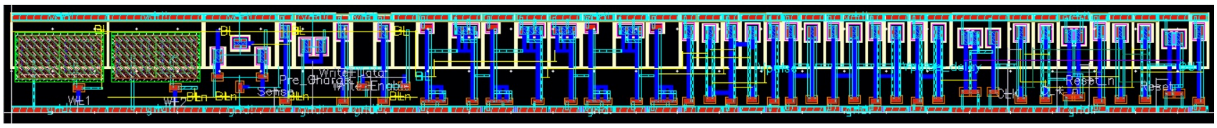


Figure 6.13: 1-bit DRAM plus Ultra-Low Power Performance Sensor layout.

Chapter 7

Conclusions

This master's thesis is another step forward in the search for a performance sensor for memories (SRAM and DRAM), to detect errors during read and write operations. The purpose is to make a sensor that can be used in the subthreshold regime.

Considering RAM memories, in particular SRAM and DRAM memories, they can be exposed to several parameter variations, causing a decrease in their performance, resulting in slower transitions, which in turn will cause slower reads and writes and can lead to errors during memory operations. The most important parameters that affect memories' performance are process (P), voltage (V), temperature (T), and aging (A), or the combination of all, PVTA. Regardless of what parameter is causing the degradation, the important aspect is to be able to measure memory performance degradation to detect it precociously and avoid error occurrence. It is therefore necessary to monitor the errors of a memory through sensors.

Process variations are caused by physical random fluctuations in transistor construction, causing small deviations in transistors' parameters like L_{eff}, number of dopants, etc. This leads to unique chips with unique timing responses.

Aging of CMOS memories affects transistor conductivity, delaying memory performance which translates into the occurrence of errors in memories over time, and this is undesirable, especially in critical systems. These aging variations are cumulative over time and are potentiated by high temperature, which lead to permanent degradation in circuits.

Power supply voltage fluctuations, due to circuits activity and power-supply distribution along the die, affects timing response of the circuit and makes the propagation delays difficult to predict. Another crucial issue regarding power in IoT applications is energy management. A wide variety of smart sensors, usually battery operated, require high energy efficiency. This is aimed at searching for ultra-low power microcontrollers and low-power memories. For this, a key variable is the minimum power supply voltage value, V_{DD}, which can ensure secure data retention and access to data (read/write operations). Using a flexible power management unit (PMU), it can be rewarding to perform dynamic voltage and frequency sizing (DVFS) to power memory matrices with minimal V_{DD} during memory access and data retention.

Also due to circuit activity, temperature hot spots appear in parts of the circuit exposed to higher activity, also affecting in an unpredictable way circuit's performance.

Some work has already been done in the search for sensors that allow monitoring the errors of a memory, such as the OCAS sensor, which detects aging in an SRAM memory caused by aging by NBTI. However, this sensor has some limitations, as it cannot be applied to DRAM memories and does not contemplate other aging effects (like the PBTI effect). Also, it concentrates on aging effects, but PVT degradations can also cause major performance degradations. But, developing specific sensors for each parameter separately is almost impossible, because other effects affect circuit with similar degradations which makes almost impossible to isolate degradations coming from various sources.

Another solution presented earlier is the performance sensor for a SRAM memory, presented in [23], that demonstrates some evolution in relation to the OCAS sensor, because it monitors memory performance, regardless of what parameter causes degradations, but still contains limitations (as is the case of being quite dependent on the synchronism with the memory and not allowing any type of calibration of the system throughout its operation). To overcome these limitations, the Scout Memory Sensor was presented in [2], which allows its use in SRAM and DRAM memories, and also allows the designer to calibrate and change the sensitivity of the sensor, making this solution more versatile and robust. However, the Scout Memory Sensor is not consistent and coherent for a subthreshold regime, as it was proven in the present thesis in chapter 4. It does not ensure error signaling when supply voltages are too low, so therefore it is necessary to find an alternative solution for this sensor that can work at lower supply voltages.

7.1 Conclusions for Scout Memory Sensor at Subthreshold Regime

The first objective of this dissertation involved the analysis of the behavior of the Scout Memory Sensor (developed by Luis Santos in [2]) to work at lower supply voltages and close to threshold voltage (V_T) (subthreshold regime). For this, several parametric simulations of the different constituent blocks of the Scout Memory Sensor were performed. Through the simulations carried out, we can conclude that the first two blocks of the sensor (transition detector and pulse detector) work well in the subthreshold regime up to supply voltages of 0.11V. However, regarding the comparator block, the simulations showed that this block has failures when working in the subthreshold regime, because this block becomes incoherent, signaling errors at higher supply voltages and, for lower supply voltages, these errors are no longer signaled. We can therefore conclude that due to the complexity of the comparator block architecture and the electronic component stack, the Scout Memory Sensor is not a reliable sensor in subthreshold regime. It is therefore necessary to find an alternative to this sensor that works at lower supply voltages, to have a sensor that can be used in IoT applications. Therefore, another objective of this dissertation is the presentation and analysis of a new sensor, the Ultra-Low Power Performance Sensor for memories.

7.2 Conclusions for the new Ultra-Low Power Performance Sensor for memories, at Subthreshold Regime

The Ultra-Low Power Performance Sensor has a different operation and simpler architecture when compared with the Scout Memory Sensor. This fact makes its use more appropriate for working at a subthreshold voltage regime.

The operation of this new sensor uses the clock that triggers write/read operations in memory to synchronize sensor operation. These write/read operations impose a sequence of transitions in the memory circuit, and the data in the memory is used by a subsequent combinatorial or sequential circuit. This propagation delays create a critical path in the memory circuitry that will impose the minimum clock period to work with the memory. To avoid errors, a safety margin is used and the minimum clock period accounts with the critical path and with the safety margin. The new sensor works by monitoring this safety margin and signaling when this safety margin is reduced, and circuit operation is in the eminence of an error. The transition detector block generates a pulse when a bit line transition occurs, due to a read/write operation in the memory. This pulse is proportional in duration to the bit line transition time. Then, a delay element adds a propagation delay to this pulse, placing it near the next clock active transition. Therefore, due to the appearance of PVTA variations (or a performance degradation), all propagation delays become higher and when the delayed pulse reaches the active edge of the clock, the sensor signals an error.

Through the simulations presented in this dissertation, we can conclude that this sensor works well for supply voltages in the subthreshold regime (namely up to 0.33 V). However, it is advised to work up to supply voltages close to 0.4 V, because below this value the sensor is no longer so cautious (decreasing the percentage of safety margin). The sensor can signal errors due to PVTA variations from the cell or from the peripheral circuits of the memory. On the other hand, this sensor can be used both in SRAM and DRAM memories, and most important for any type of memory cell architecture, because the sensor works as long as transitions occur in the bit line. Moreover, this version can be used as a local sensor, placed inside the memory, placed in all bit lines, or monitoring a small sample of cells to reduce area overhead. Also, it can be used to design a complete global sensor approach.

We can conclude that the Ultra-Low Power Performance Sensor is quite versatile. However there is one aspect that could need improvements in future work. The sensor signaling threshold, measured by the propagation delays of the sensor, is defined by designed and not possible to change during circuit lifetime. However, if the designer wants to increase sensors sensibility to bit line transitions, he can use additional unbalanced inverters in the transition detector. And, if the designer wants to increase the safety margin for the circuit in order to make the sensor activation more sensitive, or if data from memory is captured by a more complex combinatorial logic that requires a higher delay, he can use additional delay elements in the delay element block.

A good point by making everything defined by design is the fact that the sensor has a small area, and a simple functionality. This way, it can be rewarding to use the sensor in memory, even with the area overhead, because of the increased reliability.

Regarding the use of the new Ultra-Low Power Performance Sensor for memories to control a DVFS methodology, simulations have proven that the sensor can monitor effectively the reduction of the safety margin in a memory, and tunning the optimum V_{DD} / f_{CLK} pairs.

According with the optimization strategy used in the DVFS technique, the clock frequency can be adjusted whenever the supply voltage is changed, or vice-versa, so that the sensor makes the circuit to work at maximum performance, that is, for each VDD value, there is a clock frequency from which the sensor is on the eminence of indicating an error (in this work we used for normal operation 20% additional safety margin). In this way, this sensor is able to control a the operation of a DVFS technique, and its monitoring information is crucial to the DVFS controller that controls the tuning of the optimum operation. Typically, the sensor can be used in off-line tests to define a table of V_{DD} and clock frequency pairs. Then, during the online operation, the table can be constantly updated according with sensor online information, regardless the sensor is used as a local sensor, a global sensor, or both. This strategy of using both global and local sensors, represents a great advantage over previous sensors' topologies, because even if the main memory is not being used, this sensor remains updated and ready to detect errors due to PVTA variations, while circuit operation is optimized. Nevertheless, the use as a global sensor, and the DVFS controller architecture, are not part of this work, but will be addressed in future works.

7.3 Future Work

As in all the works there is always something that remains to be developed and in this section it is intended to mention works that can be carried out in the future. During this document, several aspects were already mentioned to be addressed for future work, and we resume all in this section. Future works are:

- One of these future works is to perform tests and simulations for the Ultra-Low Power Performance Sensor for writings or readings in a different SRAM memory cell than type 6T (as is the case with 8T and 10T memory cells) and verify that the sensor continues to have the same reliability and robustness.

- A new implementation of the sensor transition detector when working with SRAM memories can be done as future work, and the existing transition detector can be simplified. In resume, we can replace the fastest path of the unbalanced inverters (path 1) with a direct line connected to the XOR gate, because the SRAM is initialized to V_{DD}. In the transition detector, on path created by the unbalanced inverters is a slow path to Low-to-High transitions in the bit line, while the other path is a slow path to High-to-Low transitions. However, as the SRAM is used with a pre-charge value of V_{DD}, sensor will only be activated by High-to-Low transitions, therefore we can replace the slow Low-to-High path (which is has a fast High-to-Low transition) by a fast path, i.e., a direct line. This new model of the transition detector was not studied because the goal was to maintain a simple and unique structure for the transition detector, by reusing this block from previous work, so that it could be used in all memory types (SRAM and DRAM).

- Another future work is the development of a DVFS controller that allows the sensor to operate at low voltages, namely under subthreshold regime, to establish the V_{DD} / f_{CLK} optimum pairs.
- On the other hand, it is necessary to implement the sensor as a global sensor, with its circuitry, including a dummy memory, sensors, and a controller.
- An important future work is to change the delay element block in order to include a sensibility control to the sensor, i.e., to allow the propagation delay of the delay element to be changed during online operation.
- Finally, a work to be carried out in the future is the implementation of the sensor in a chip, with the aim of being possible to validate in silicon all sensor functionalities described.

Bibliography

[1] Hugo Fernandes da Silva Santos. Aging sensor for cmos memory cells. Master thesis on engenharia elétrica e eletrónica, especialization in tecnologias de informação e telecomunicações, Instituto Superior de Engenharia da Universidade do Algarve, September 2015.

[2] Luís Mário Braz dos Santos. Performance sensor for cmos memory cells. Master thesis on engenharia elétrica e eletrónica, especialization in tecnologias de informação e telecomunicações, Instituto Superior de Engenharia da Universidade do Algarve, October 2020.

[3] S. Marusic J. Gubbi, R. Buyya and M. Palaniswami. Internet of things (iot): A vision, architectural elements, and future directions. *Future Generat. Comput. Syst.*, 29(7):1645–1660, 2013.

[4] W. Saad T. Park, N. Abuzainab. Learning how to communicate in the internet of things: Finite resources and heterogeneity. *IEEE Access: Optimization for Emerging Wireless Networks: IoT, 5G and Smart Grid Communication Networks, Special Session*, 4(7):7063–7073, November 2016.

[5] P. Han Joo Chong H. Yu Shwe, T. King Jet. An iot-oriented data storage framework in smart city applications. *2016 International Conference on Information and Communication Technology Convergence (ICTC)*, pages 106–108, 2016.

[6] S K Alamgir Hossain A. Mohon Ghosh, D. Halder. Remote health monitoring system through iot. *2016 5th International Conference on Informatics, Electronics and Vision (ICIEV)*, pages 921–926, 2016.

[7] Tadaaki Yamauchi; Hiroyuki Kondo; Koji Nii. Automotive low power technology for iot society. *2015 Symposium on VLSI Circuits (VLSI Circuits)*, pages T80–T81, 2015.

[8] H. He et al. The security challenges in the iot enabled cyber-physical systems and opportunities for evolutionary computing other computational intelligence. *2016 IEEE Congress on Evolutionary Computation (CEC)*, pages 1015–1021, 2016.

[9] T. Xu; J. B. Wendt; M. Potkonjak. Security of iot systems: Design challenges and opportunities. *Proc. IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, pages 417–423, 2014.

[10] K. Kaur K. Kaur. A study of power management techniques for internet of things (iot). *Proc. Int.*

Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), pages 1781–1785, 2016.

[11] E. Sanchez-Sinencio S. Carreon-Bautista, L. Huang. An autonomous energy harvesting power management unit with digital regulation for iot applications. *IEEE Journal of Solid-State Circuits*, 51(6):1457–1474, 2016.

[12] Chingwei Yeh De-Shiuan Chiou, Shih-Hsin Chen. Timing driven power gating. *Proceedings of the 43rd annual conference on Design automation*, pages 121–124, 2006.

[13] J. Rodríguez-Andina L. Piccoli F. Vargas M. Santos I. Teixeira J. Semião, M. Irago and J. Teixeira. Signal integrity enhancement in digital circuits. *IEEE Design and Test of Computers*, 25(5):452–461, October 2008. DOI: http://dx.doi.org/10.1109/MDT.2008.146.

[14] S. Bhardwaj-R. Vattikonda S. Vrudhula F. Liu W. Wang, S. Yang and Y. Cao. The impact of nbti on the performance of combinational and sequential circuits. *in Proc. of the ACM/IEEE Design Automation Conference*, 8(4):364–369, June 2007. DOI:http://dx.doi.org/10.1109/DAC.2007.375188.

[15] T. Kim and Z. Kong. Impact analysis of nbti/pbti on sram vmin and design techniques for improved sram vmin. *in Journal of Semiconductor Technology and Science*, 13(2):87–97, April 2013. DOI:http://dx.doi.org/10.5573/JSTS.2013.13.2.87.

[16] L. Bolzani A. Ceratti, T. Copetti and F. Vargas. On-chip aging sensor to monitor nbti effect in nano-scale sram. *in Proc. of the 2012 IEEE 15th International Symposium on Design and Diagnostics of Electronic Circuits and Systems, DDECS*, 18(20):354–359, April 2012. DOI:http://dx.doi.org/10.1109/DDECS.2012.6219087.

[17] U. Shirode A. Gadhe. Read stability and write ability analysis of different sram cell structures. *International Journal of Engineering Research and Applications (IJERA)*, 3(1):1073–1078, January 2013.

[18] M. Sachdev M. Sharifkhani. Sram cell stability: A dynamic perspective. *IEEE Journal of Solid-State Circuits*, 44(2), February 2009.

[19] D. Saraiva-C. Leong M. Santos I. Teixeira J. Semiao, A. Romao and P. Teixeira. Performance sensor for tolerance and predictive detection of delay-faults. *Proc. of the Int. Symposium on Defect and Fault Tolerance in VLSI and Nanotechnology Systems (DFT)*, 1(3):34–39, October 2014. DOI:http://dx.doi.org/10.1109/DFT.2014.6962092.

[20] J. Vazquez-V. Champac M. Santos I. Teixeira C. Martins, J. Semião and P. Teixeira. Adaptive error–prediction flip–flop for performance failure prediction with aging sensors. *in IEEE 29th– VLSI Test Symposium (VTS)*, 1(5):203–208, May 2011. DOI:http://dx.doi.org/10.1109/VTS.2011.5783784.

[21] M.J. Moure-J.J. Rodríguez-Andina J. Semião F. Vargas I.C. Teixeira J.P. Teixeira M. Valdés, J. Freijedo. Design and validation of configurable on-line aging sensors in nanometer-scale fpgas. *IEEE*

Transactions on Nanotechnology, Special Issue on ”Defect Fault Tolerance in VLSI and Nanotechnology Systems”, 12(4):508–517, July 2013.

[22] C. Leong-M. Santos-I. Teixeira J. Teixeira J. Semião, R. Cabral. Dynamic voltage scaling with fault-tolerance for lifetime operation. *in the 4th. Workshop on Manufacturable and Dependable Multicore Architectures at Nanoscale (MEDIAN’15) / DATE’2015 Workshop W06*, March 2015.

[23] R. Cabral-A. Romão-M. Santos I.C. Teixeira J.P. Teixeira J. Semião, H. Santos. Aging and performance sensor for sram. *Proc. of the 31th. Design of Circuits and Integrated Systems Conference (DCIS 2016)*, 23(25), November 2016.

[24] R. Cabral-M. B. Santos J. Semião, H. Santos and P. Teixeira. Pvta-aware performance sram sensor for iot applications. *Proceedings of the 2nd INternational CongRess on Engineering and Sustainability in the XXI cEntury – INCREaSE 2019*, 2019 pages = 337–353, month = oct, note = DOI 10.1007/978-3-030-30938-1.

[25] Luís Santos Jorge Semião and Marcelino Santos. Sram performance sensor. *Proceedings of the Proceedings of the XXXVI edition of the Design of Circuits and Integrated Systems Conference (DCIS 2021)*, November 2021.

[26] Marcelino B. Santos Jorge Semiao, Luis Santos. Dram performance sensor. *Proceedings of the 24th International Conference On Human-Computer Interaction*, June 2022.

[27] T. Grasser. *Bias Temperature Instability for Devices and Circuits*. Springer, 2013.

[28] M.C.; Moon J.E.; Ko P.K.; Hu C. Chung, J.E.; Jeng. Performance and reliability design issues for deep-submicrometer mosfets. *IEEE Trans. Electron Devices*, 7(38):545–554, 1991.

[29] B.; Narayanan V.; Paruchuri V. Cartier, E.; Linder. Fundamental understanding and optimization of pbti in nfets with sio2/hfo2 gate stack. *In Proceedings of the International IEEE Electron Devices Meeting (IEDM’06)*, pages 1–4, December 2006.

[30] Dram size calculation. *SystemVerilog.io web page*, July 2022. https://www.systemverilog.io/ddr4-basicssize-calculation.

[31] C.M. Jeppson, K.O.; Svensson. Negative bias stress of mos devices at high electric fields and degradation of mnos devices. *J. Appl. Phys.*, 48(16):2004–2014, 1977.

[32] D.K. Schroder. Negative bias temperature instability: What do we understand? *Microelectron. Reliab.*, 47(11):841–852, 2007.

[33] K.; Tse K.Y. Robertson, J.; Xiong. Importance of oxygen vacancies in high k gate dielectrics. *In Proceedings of the IEEE International Conference on Integrated Circuit Design and Technology (ICICDT’07)*, 2007 pages = 1–4, month = may,.

[34] K.; Chaudhary V.; Goel N.; De S.; Pandey R.K.; Murali K.V.R.M.; Mahapatra S. Mukhopadhyay, S.; Joshi. Trap generation in il and hk layers during bti/tddb stress in scaled hkmg n and p mosfets. *In Proceedings of the 2014 IEEE International Reliability Physics Symposium*, pages GD.3.1–GD.3.11, June 2014.

[35] S.; Goel N.; Nanware N.; Mahapatra S. Joshi, K.; Mukhopadhyay. A detailed study of gate insulator process dependence of nbti using a compact model. *IEEE Trans. Electron Devices*, 61(23):408–415, 2014.

[36] S.; Goel N.; Mahapatra S. Joshi, K.; Mukhopadhyay. A consistent physical framework for n and p bti in hkmg mosfets. *In Proceedings of the 2012 IEEE International Reliability Physics Symposium (IRPS)*, pages 5A.3.1–5A.3.10, April 2012.

[37] K. Roy K. Kang, H. Kufluoglu and M. A. Alam. Impact of negative bias temperature instability in nanoscale sram array: Modeling and analysis. *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, 26(10):1770–1781, October 2007.

[38] K. Roy K. Kang, S. P. Park and M. A. Alam. Estimation of statistical variation in temporal nbti degradation and its impact on lifetime circuit performance. *in Proc. IEEE/ACMInt. Conf. Comput.–Aided Des.*, pages 730–734, November 2007.

[39] A. Haggag et al. Understanding sram high-temperature-operatinglife nbti: Statistics and permanent vs recoverable damage. *in Proc. 45th Annu. IEEE Int. Rel. Phys. Symp.*, pages 452–456, April 2007.

[40] K. Kim S. V. Kumar and S. S. Sapatnekar. Impact of nbti on sram read stability and design for reliability. *in Proc. 7th Int. Symp. Quality Electron. Des. (ISQED)*, page 6, March 2006.

[41] V. Huard et al. Nbti degradation: From transistor to sram arrays. *in Proc. IEEE Int. Rel. Phys. Symp.*, pages 289–300, April 2008.

[42] Chandrakasan A. Calhoun, B.H. Characterizing and modeling minimum energy operation for subthreshold circuits. *In: International symposium on low power electronics and design*, pages 90–95, 2004.

[43] Wang Alice Chandrakasan Anantha Calhoun, Benton H. Modeling and sizing for minimum energy operation in subthreshold circuits. *IEEE J. Solid-State Circuits*, 40(9):1778–1786, September 2005.

[44] Wang A. Chandrakasan A. Calhoun, B.H. Device sizing for minimum operation in subthreshold circuits. *In: Custom integrated circuits conference*, 2004.

[45] Wang A. Verma N. Chandrakasan A. Calhoun, B.H. Sub-threshold design: the challenges of minimizing circuit energy. *In: Proceedings of the international symposium on low power electronics and design (ISLPED)*, October 2006.

[46] Bhargav S. Moore C. Martin A.J. Keller, S. Reliable minimum energy cmos circuit design. *In: Vari'11: 2nd european workshop on CMOS variability*, 2011.

[47] H.J. Yoo. Dual vt self-timed cmos logic for low subthreshold current multigigabit synchronous dram. *IEEE Trans. Circ Sys-II: Analog Digital Signal Process*, 45(9):1263–1271, September 1998.

[48] Seok M. Sylvester D. Blaauw D. Hanson, S. Nanometer device scaling in subthreshold logic and sram. *IEEE Trans. Electron Devices*, 55:175–185, 2008.

[49] Boon C.C. Do M.A. Yeo K.S. Cabuk A. Do, A.V. A subthreshold low-noise amplifier optimized for ultra-low-power applications in the ism band. *IEEE Trans. Microw. Theory Tech.*, 56(2):286–292, February 2008.

[50] Palumbo G. Criscione M. Cutri F. Giustolisi, G. A low-voltage low-power voltage reference based on subthreshold mosfets. *IEEE J. Solid-State Circuits*, 38(1):151–154, January 2003.

[51] Roy K. Kim, J.J. Double gate mosfet subthreshold circuit for ultralow power applications. *IEEE Trans. Electron Devices*, 51(9):1468–1474, September 2004.

[52] Takagi S. Numata, T. Device design for subthreshold slope and threshold voltage control in sub-100-nm fully depleted soi mosfets. *IEEE Trans. Electron Devices*, 51(12):2161–2167, December 2004.

[53] MZ et al. Li. Sub-threshold standard cell library design for ultra-low power biomedical applications. *In: Engineering in medicine and biology society (EMBC) 2013 35th annual international conference of the IEEE*, page 1454, 2013.

[54] Eappen G. Sahu, A. Sub-threshold logic and standard cell library. int. j. innovative res. sci. *Eng. Technol.*, 3(1), January 2014.

[55] Stojanovic V. Nikolic B. Horowitz M.A. Brodersen R.W. Markovic, D. Methods for true energy-performance optimization. *IEEE J. Solid-State Circuits*, 39(8), August 2004.